Heidi Iivonen

Finding the key genes in plasma cell differentiation and their effect on antibody secretion in *Saccharomyces cerevisiae*

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 4.11.2015

Thesis supervisor:

Prof. Alexander Frey

Thesis advisor:

M.Sc. Essi Koskela

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Heidi Iivonen

Title: Finding the key genes in plasma cell differentiation and their effect on antibody secretion in *Saccharomyces cerevisiae*

Date: 4.11.2015 Language: English Number of pages: 7+97

Department of Biotechnology and Chemical Technology

Professorship: Molecular Biotechnology Code: KEM-70

Supervisor: Prof. Alexander Frey

Advisor: M.Sc. Essi Koskela

Baker's yeast *Saccharomyces cerevisiae* is putatively an attractive host for producing antibodies. The production yields are low, however, and therefore the ongoing research is focused on improving the productivity of the yeast. In contrast, plasma cells are very efficient in producing antibodies, thus the objective of this thesis was to utilize information on plasma cells gained from enrichment analysis when selecting genes for experiments in *S. cerevisiae*.

The enrichment analysis was conducted on nine data subsets from four data sets containing transcriptomic data from naïve B cells and plasma cells. The analysis yielded total of 142 enriched terms, out of which seven were chosen for closer inspection to see which genes comprise them. Four genes, *SBH1*, *GPD2*, *TRX1* and *TRX2*, were selected to be overexpressed and knocked out in antibody-secreting yeast. The results on the secretion of antibody were analysed by ELISA.

In the *in vivo* testing, the overexpression of *TRX2* increased most the secretion of antibody. In contrast, the overexpression of *TRX1* resulted in lower antibody concentrations. The overexpression of *SBH1* had conflicting results under different growth conditions, and its knockout didn't affect the secretion of antibody. The knockout of *GPD2* increased the relative antibody secretion.

Overall, the results of this thesis show that combining enrichment analysis and experiments *in vivo* is a feasible way to enhance the production of antibodies in yeast.

Keywords: *Saccharomyces cerevisiae*, gene enrichment analysis, plasma cell, antibody

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Heidi Iivonen

Työn nimi: Plasmasolujen tärkeimpien geenien löytäminen ja niiden vaikutukset vasta-ainetuottoon *Saccharomyces cerevisiae* –hiivassa

Päivämäärä: 4.11.2015 Kieli: Englanti Sivumäärä: 7+97

Biotekniikan ja kemian tekniikan laitos

Professuuri: Molekyylibioteknologia Koodi: KEM-70

Työn valvoja: Prof. Alexander Frey

Työn ohjaaja: FM Essi Koskela

Saccharomyces cerevisiae –hiiva on lupaava organismi vasta-aineiden tuotantoon. Kuitenkin toistaiseksi saanto on vähäistä, joten meneillään oleva tutkimus keskittyy tuotannon parantamiseen. Plasmasolut sen sijaan ovat hyvin tehokkaita vasta-aineiden tuottajia, joten tämän työn tarkoituksena oli hyödyntää plasmasolujen geenien rikastumisanalyysistä saatua tietoa valittaessa geenejä, mitä muokata vasta-aineita tuottavassa *S. cerevisiae* –hiivassa.

Rikastusmisanalyysissä tutkittiin transkriptomiikkadataa neljästä eri tutkimuksesta. Data koostui B-solujen ja plasmasolujen transkriptioista, ja se oli jaettu yhdeksään alaosioon. Tuloksena oli yhteensä 142 rikastunuttua biologista termiä, joista seitsemästä tutkittiin tarkemmin, mistä geeneistä ne koostuivat. Rikastumisanalyysissä löydetyistä geeneistä neljä, *SBH1, GPD2, TRX1* ja *TRX2*, valittiin vaimennettavaksi ja yliekspressoitaviksi vasta-aineita tuottavassa hiivassa. Vasta-ainetuoton tulokset analysoitiin ELISAlla.

In vivo – kokeissa *TRX2*-geenin yliekspressio paransi parhaiten vasta-ainetuottoa, mutta kontrastina sille *TRX1*-geenin yliekspressio huononsi vasta-aineen tuotantoa. *SBH1*-geenin yliekspressio antoi ristiriitaisia tuloksia eri kasvatusolosuhteissa, ja sen vaimentaminen ei vaikuttanut vasta-ainetuottoon. *GPD2*-geenin vaimentaminen paransi suhteellista vasta-ainetuotantoa.

Tämän työn tulokset näyttävät, että rikastumisanalyysin ja laboratoriokokeiden yhdistäminen on yksi strategia hiivan vasta-ainetuoton parantamiseen.

Avainsanat: *Saccharomyces cerevisiae*, vasta-aine, geenien rikastuminen, plasmasolu

Preface

First of all, I would like to thank Professor Alexander Frey for both providing the opportunity for this thesis and for his supervision and insightful feedback on this thesis. I thank my instructor Essi Koskela for her ideas for the thesis subject, guidance and valuable feedback on this thesis. I want to thank the whole research group with their help especially in the laboratory, and for making me feel welcome from the very first coffee break.

I want to thank my parents, Pekka and Kirsti, for always believing in me, supporting me, and emphasizing the importance of education. I thank my husband Mika for his endless love and support. Finally, I want to dedicate this thesis for our beloved daughter Lilja, I hope you will never lose your curiosity.

Espoo, 4.11.2015.

Heidi A. Iivonen

Contents

Symbols and abbreviations

Ab	Antibody
BP	Biological process
ELISA	Enzyme-linked immunosorbent assay
ER	Endoplamic reticulum
ERAD	ER-associated degradation
GEA	Gene enrichment analysis
GO	Gene ontology
GPD2	Glycerol-3-phosphate dehydrogenase 2 (mitochondrial)
Ig	Immunoglobulin
IgG	Immunoglobulin G
Mab	Monoclonal antibody
MF	Molecular function
PC	Plasma cell
SBH1	Sec61 beta subunit
scFv	single-chain variable fragment
TRX1	Thioredoxin 1
TRX2	Thioredoxin 2
UPR	Unfolded protein response

1 Introduction

Antibody fragments and monoclonal antibodies (mAbs) count currently as the most important and fastest growing class of biopharmaceutical products, i.e. products which are nucleic-acid, cell, or tissue-based, or recombinant therapeutic proteins.(Spadiut et al., 2014; Walsh, 2010). In 2013, the global sales of biopharmaceuticals was estimated to reach US$140 billion per year, which is larger than the gross domestic product (GDP) of three quarters of countries listed in the GDP database of the World Bank (Walsh, 2014).

Several biopharmaceuticals are produced in the yeast *Saccharomyces cerevisiae*, and it was the first yeast used in the recombinant protein production. Yeast is an attractive host for the production of antibody fragments and mAbs, due to its status as "generally recognized as safe" (GRAS), well-known biochemical and molecular properties plus manipulation techniques, the ease of cultivation, and the stability of its expression system. However, its secretory capacity and productivity is still low. The efficient production of antibodies is hindered by inefficient trafficking in the secretory pathway and misfolding in the endoplasmic reticulum (ER). Current research is focused on improving the production yield of *S. cerevisiae* (Berlec and Strukelj, 2013; Spadiut et al., 2014).

One possible way to improve the secretion of antibodies in *S. cerevisiae* is to look into nature and see how the antibody-secreting plasma cells function. This can be done by enrichment analysis, which is a high-throughput method for identifying which biological functions play a key role in a given biological phenomenon. If a certain process is relevant in plasma cells, it is likely to be found as enriched by statistical methods when going through transcriptomics data (Huang et al., 2009). The genes involved in this process can then be modified in antibody-secreting *S. cerevisiae* in order to improve its production yield.

The target of this thesis was to find out how the enriched genes of plasma cells affected the antibody production of *Saccharomyces cerevisiae*. The aim of the enrichment analysis was to find out which biological processes and molecular functions have a significant role in the functioning of plasma cells, and which genes contribute to these enriched functions. Four of the enriched genes were selected for further testing. They were both overexpressed and knocked out in antibody-secreting *S. cerevisiae* in order to see whether they would enhance antibody secretion also in yeast.

This thesis consists of a literature part and experimental part. The literature part covers an overview of three main topics regarding this thesis. Plasma cell development and properties of antibodies are introduced. Antibody production in *S. cerevisiae* and its challenges are presented. Finally, transcriptomics and enrichment analysis are described. In the experimental part, the used materials and methods are described. The results are divided into two parts, the results of the enrichment analysis are described first, and the second part discusses the results of the overexpression and knockout studies on *S. cerevisiae*.

2 Plasma cells

Immunity means the ability of the body to defend itself against specific pathogens such as bacteria, viruses, toxins, and foreign tissue. Antigens are substances recognized as foreign, and they provoke the immune response. The immune system is divided into innate and adaptive immune system. The innate immune system is responsible for rapid response to an antigen, and the adaptive system forms a specific response to a previously encountered pathogen. Both immune systems function by first recognizing antigen and then working on eliminating or neutralizing the invader. The innate immune system recognizes structural elements, like certain glycolipids, that are present in many pathogens but absent from the host organism. The adaptive immune system is able to form over 10^8 distinct antibodies which recognize specific antigens. Due to the specificity of the adaptive immune system, the second encounter with an antigen usually elicits more rapid and vigorous response (Berg et al., 2007; Tortora and Derrickson, 2006).

In the adaptive immune system, there are two parallel and interrelated systems: humoral and cellular immune responses. In the cellular response, cytotoxic T lymphocytes (killer T cells) kill cells invaded by the pathogen. In the humoral response, antibodies travel usually in lymph and blood to the antigen invasion site, where they function as recognition elements which bind to foreign molecules and act as markers that signal foreign invasion (Berg et al., 2007; Tortora and Derrickson, 2006).

Plasma cells are antibody-secreting cells that play a key role in the humoral immunity response. They develop from B cells, which are activated by antigens, after which they proliferate and differentiate into plasma cells or activated T-cells. After exposure to an antigen, plasma cells secrete antibodies until the cells themselves die (Berg et al., 2007; Tortora and Derrickson, 2006). Plasma cells secrete antibodies but do not proliferate, unlike some fractions of the proliferating B cells, called plasmablasts, which secrete measurable amounts of antibody (Oracki et al., 2010).

2.1 Antibodies

The function of antibodies is to bind to antigens, and to be a facilitator of their removal from the body. Generally an antibody recognizes only a certain small part, an epitope, of an antigen, for instance certain amino acids of a viral coat protein (Janeway et al., 2001).

Antibodies are also called immunoglobulins (Igs), since they belong to a group of glycoproteins called globulins. Most antibodies comprise four polypeptide chains, of which two are called heavy (H) chains, and the other two light (L) chains. Both heavy and light chains are identical with their respective counterpart. Heavy chains consist of about 450 amino acids, and they have short carbohydrate chains attached to them. Light chains consist of about 220 amino acids, and each one is connected to a heavy chain by a disulfide bond (Tortora and Derrickson, 2006). There are two types of light chains, λ and κ (Berg et al., 2007). The heavy chains of IgG are also connected to each other via two disulphide bonds that reside in the midregion of the heavy chains. This area called hinge region is very flexible, and thus the antibody can

be shaped either T or Y like by bending the hinge region. The antigen-binding sites, that recognize and attach to specific antigens, are located at the tips of the heavy and light chains called the variable (V) regions. The bending of the hinge region allows the antigen-binding sites to attach to two identical epitopes that are apart from each other. The region adjacent to the variable region is called the constant (C) region, and the structure of the heavy chains in this region is used as a basis for distinguishing the five different antibody classes (Tortora and Derrickson, 2006). The structure of an antibody molecule is illustrated in picture 1.

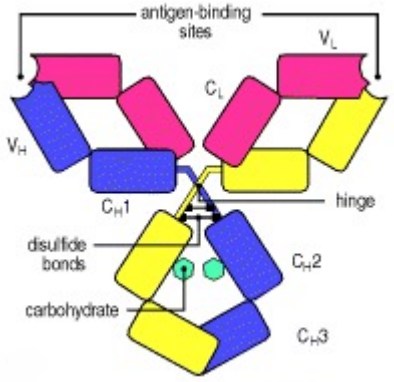

Figure 1: Antibody molecule structure. Pink parts are light chains; yellow and purple parts heavy chains. V_H and V_L refer to respective variable regions of the heavy and light chains, and parts C_H1-3 make up the constant region with their identical counterparts (Janeway et al., 2001).

The domains of an antibody have many common sequence features, and they adopt a common structure, the immunoglobulin fold, which is also found in many other proteins playing key functions in the immune system. The immunoglobulin fold consists of a pair of β sheets, built of antiparallel β strands surrounding a central hydrophobic core. The sheets are connected by a disulphide bridge. Three hypervariable loops, called that because the hypervariable sequence of an antibody is located there, forming a potential binding surface reside at the surface near the N terminus. The C and N terminuses of immunoglobulin fold are at the opposite ends of the immunoglobulin fold, allowing them to form chains, like the the light and heavy chains of an antibody, by being strung together (Berg et al., 2007).

The five different antibody classes are IgM, IgD, IgG, IgA, and IgE, and they have heavy chain names corresponding to their respective lower-case Greek letter ($\mu, \delta, \gamma, \alpha$, and ϵ, respectively) (Janeway et al., 2001). IgG is the most abundant one, about 80% of all antibodies in the blood belong to this class. In addition to blood, it is also found in lymph and the intestines. It enhances phagocytosis, neutralizes toxins, and triggers the complement system, which is a defensive system made of a set of proteins that destroy microbes collectively. IgA accounts for about 10-15% of antibodies in blood, however it is mainly found in sweat, saliva, tears, mucus, breast milk, and gastrointestinal secretions. It provides localized protection against viruses and bacteria on mucous membranes. IgM is present in blood and lymph, it

comprises about 5-10% of antibodies in blood. It activates the complement system, and causes agglutination and lysis of microbes. It is the first one to be secreted by plasma cells, thus its' presence is an indicator of a recent invasion. IgD is involved in the activation of B cells, and it is found mainly on their surfaces. It accounts only for about 0.2% of antibodies in blood. IgE resides in mast cells and basophils, and it is involved in allergic reactions and in providing protection against parasitic worms. Less than 0.1% of antibodies in blood are IgE. IgG, IgD, and IgE occur only as monomers, IgA can also occur as dimers, and IgM occur in pentamers (Tortora and Derrickson, 2006).

2.2 Development of plasma cells

Plasma cell development begins by B cell activation by the antigen, which occurs when an antigen binds to B cell receptors (BCRs) on the surface of the B cell. BCRs are monomers of IgM, and each B cell has its own specific type of IgM expressed on its surface. The plasma cells developing from a certain B cell will all secrete antibodies chemically similar to the IgM of the B cell. Different antigens stimulate different B cells to develop into plasma cells, and thus the body is able to form an immune response against a large variety of pathogens (Berg et al., 2007; Tortora and Derrickson, 2006).

The activated B cells process the antigen by taking it into the cell and breaking it down into peptide fragments. These fragments are combined with major histocompatibility complex class II (MHC-II) proteins, which function mainly as antigen-presenting molecules. MHC-II proteins present the antigens to CD4 cell-surface proteins of helper T cells, which deliver the costimulation needed for B cell proliferation and differentiation(Berg et al., 2007; Madigan et al., 2009; Tortora and Derrickson, 2006). A single MHC protein can bind to peptides sharing a unique peptide motif, for instance a phenylalanine at position 5 and a leucine at position 8. As long as the peptide has the correct residues, the MHC protein can bind to it. Thus all antigens will have at least few peptides forming a peptide motif that will be bound by MHC proteins (Madigan et al., 2009).

2.2.1 Germinal center reaction

The process of selecting which cells develop and reproduce is critical to the immune system. The immune response is based on selecting the cells expressing molecules that are effective against a particular foreign invader (Berg et al., 2007). This selection is done in the germinal centers, which are the sites where B cells develop into high-affinity antibody-secreting plasma cells and memory B cells (Victora and Nussenzweig, 2012).

Germinal centers (GCs) are areas of high mitotic activity in lymph nodes. They can be divided into two zones, dark and light one. The dark zone appears dark in light microscopy, as it is almost entirely made of B cells having a high nucleus-to-cytoplasm ratio. The light zone consists of B cells interspersed among a network of follicular dendritic cells that make it appear lighter. It also contains T cells expressing surface

proteins CD8 and CD4 on their surface. The suggested model for germinal center reaction by Victora and Nussenzweig (2012) is illustrated in figure 2 (Victora and Nussenzweig, 2012).

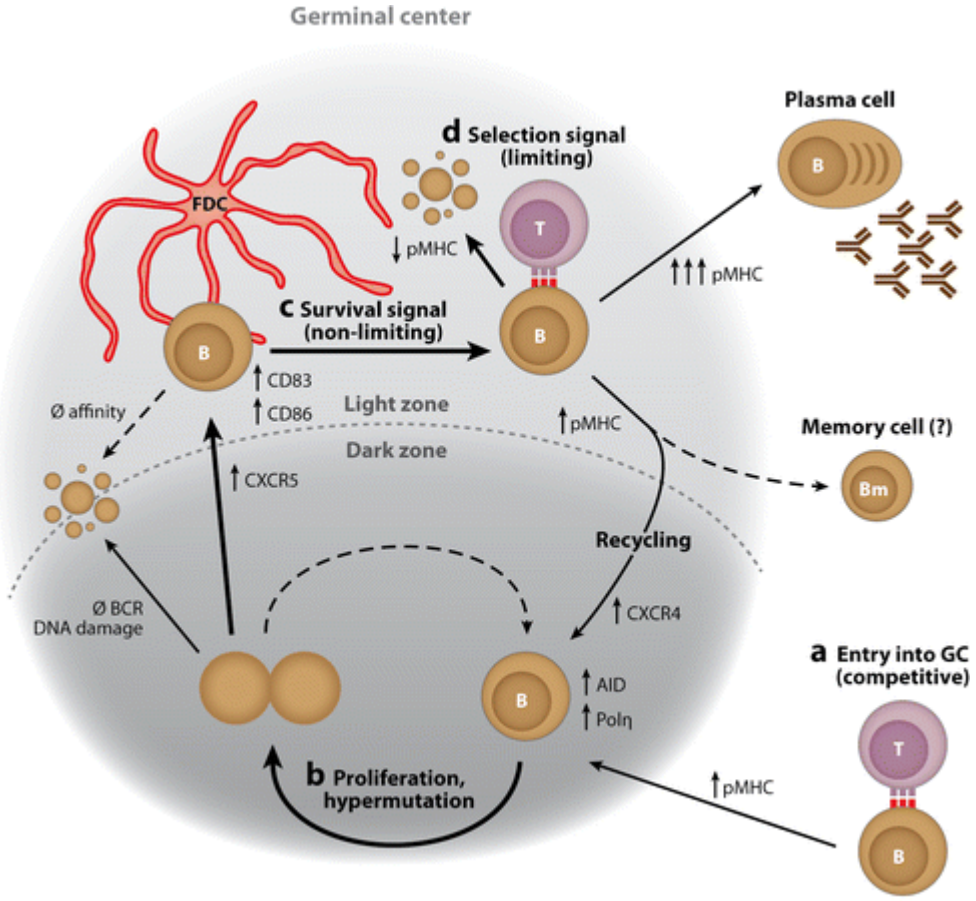

Figure 2: Proposed model of GC reaction by Victora and Nussenzweig (2012). a) Before entering the reaction, B cells compete for a limiting number of T cells, and only the ones that have the highest affinity are selected for the GC reaction. b) In the dark zone, cells proliferate and undergo somatic hypermutation, with the aid of AID and Polη enzymes. High expression of chemokine receptor CXCR4 maintains these cells in the dark zone. Cells with impaired BCR expression (∅ BCR) die due to the lack of a BCR signal. After one or more cycles of division and somatic hypermutation, B cells migrate towards the light zone. This process involves the upregulation of chemokine receptor CXCR5. c) B cells interact with antigen in immune complexes on follicular dendritic cells (FCDs). At this stage, B cells that have very low or zero affinity for antigen may also die due to the missing BCR signal. At this point, antigen signals are not limiting, and all B cells that transit to the light zone upregulate CD83 and CD86. The level of peptide-MHC (pMHC, red rectangles) on the B cell surface is determined by BCR affinity. d) B cells present antigen to T helper cells, which amount is limited. B cells that have higher pMHC are preferred in this competitive process. B cells successfully interacting with T cells have then three different fates. They may re-enter the dark zone with upregulation of CXCR4, and undergo further cycles of proliferation and hypermutation. They can become memory B cells. They can also exit the GC and become plasma cells. This third option is likely for cells that have high affinity or pMHC density. In the picture the width of an arrow represents the proportion of cells moving towards that path. Dashed arrows represent events for which data was found to be inconclusive (Victora and Nussenzweig, 2012).

Before returning to the light zone, B cells undergo a variable number of divisions in the dark zone. It has been shown in a study by Gitlin et al. (2014) that the increased cell division is associated with increased somatic hypermutation and higher immunoglobulin affinity. Therefore the high-affinity germinal center B cells are able to outcompete their low-affinity counterparts by diving greater number of times and capturing more antigen during each cycle (Gitlin et al., 2014).

2.2.2 Factors promoting plasma cell fate

B-lymphocyte induced maturation protein (Blimp1), transcription factor XBP1 (X-box binding protein 1), and interferon regulatory factor-4 (IRF4) are three main factors promoting the plasma cell fate in the germinal center reaction (Nutt et al., 2011).

IRF4 has been found in a genome-wide expression analysis by Sciammas et al. (2006) to regulate the induction of the entire Blimp1-dependent plasma cell development. B cell lacking IRF4 were severely compromised for class-switch recombination and plasma cell generation, because of their failure to induce properly the expression of genes encoding AID (activation-induced cytidine deaminase) and Blimp1, respectively. Their results suggest that the graded expression of IRF4 promotes the transition of a germinal center B cell to plasma cell (Sciammas et al., 2006). Similar results were obtained by Klein et al. (2006), who generated mice in which the gene encoding IRF4 could be conditionally deleted in germinal center B cells and the fate of the cells having the deletion could be monitored *in vivo*. They found that the mice having the deletion lacked post-germinal center plasma cells and were unable to differentiate memory B cells into plasma cells. It was found that plasma cell differentiation needed IRF4 and Blimp1, both acting upstream of XBP1. It was also found that class-switch recombination was dependent of IRF4 (Klein et al., 2006).

Blimp1 is a zinc-finger containing transcription factor regulating a vast amount of genes inducing the differentiation of multiple cell types. It is required for B cell development into fully mature plasma cells (Shapiro-Shelef et al., 2003). It has been found by Lin et al. (2002) to repress the gene Pax-5 encoding the B-cell lineage-specific activator protein (BSAP). BSAP is critical for the cells committing into B-cell lineage. Lin et al. (2002) found that Blimp1 bound a site on the Pax-5 promoter, and repressed the promoter in a binding-site-dependent manner. This repression was required for the plasma cell development, and it was concluded that the repression of Pax-5 by Blimp1 is critical for the development of plasma cells (Lin et al., 2002). Shaffer et al. (2002) found that the introduction of Blimp1 into B cells blocked a large set of genes important for B cell receptor signalling, germinal center B cell function and proliferation. At the same time it allowed the expression of important plasma cell genes, including the transcription factor XBP1. (Shaffer et al., 2002) In a study by Shaffer et al. (2004), it was found that Blimp1-deficient B cells did not upregulate Xbp1 or most other genes specific to plasma cells. Blimp1 was found to repress genes which in turn encode the repressors of plasma cell genes. It was also required for the maximal induction of immunoglobulin genes (Shaffer et al., 2004). It can be concluded that Blimp1 functions as a master regulator of

plasma cell differentiation by enhancing immunoglobulin expression, shutting down unnecessary cellular processes, and facilitating the induction of XBP1 (Brewer and Hendershot, 2005).

XBP1 is required for the terminal differentiation of plasma cells (Reimold et al., 2001). Only the spliced form of XBP1, XBP1(S), is able to efficiently activate the unfolded protein response. It is spliced by serine/threonine-protein kinase/endoribonuclease IRE1. Activating transcription factor 6 (ATF6) induces the XBP1 and ER chaperone genes via direct binding to ER stress-responsive element (ERSE) (Yoshida et al., 2001). It has been found to coordinate diverse changes in the cellular functioning and structure leading to secretory cells in a gene expression profiling study by Shaffer et al. (2004). In the study its expression induced a vast spectrum of genes involved in the secretory pathway and the physical expansion of the ER. They also found that the expression of XBP1 increased cell size, lysosome content, mitochondrial function and mass, ribosome numbers, and total protein synthesis. Its overexpression led to the overexpression of genes encoding components in the secretory pathway such as proteins targeting and translocating nascent polypeptides into the ER, chaperones and their cofactors promoting protein folding, oxidoreductases regulating protein folding, glycosylation enzymes, and vesicular trafficking regulators. In the same study it was concluded that XBP1 acted downstream of Blimp1, since almost all of the genes requiring XBP1 required also Blimp1, but a number of plasma cell genes required only Blimp1 (Shaffer et al., 2004).

2.2.3 Unfolded protein response

Unfolded protein response (UPR) is a multidimensional signalling pathway regulating the expression of ER resident proteins and lipids, translation, the cell cycle, and apoptosis. The components of UPR also regulate cell differentiation and are involved in innate immunity. UPR is triggered by the accumulation of unfolded proteins in the ER (Brewer and Hendershot, 2005; Hetz, 2012; Masciarelli and Sitia, 2008). It aims to reduce the stress inflicted into the ER from the unfolded protein load by mechanisms like the expansion of the membrane of ER, the attenuation of protein uptake into the ER, and the synthetisation of molecules involved in protein folding and quality control. If the homeostasis is not restored, apoptosis is triggered by the UPR (Hetz, 2012).

UPR has three branches that are initiated by different sensors in response to ER stress. These sensors are inositol-requiring protein 1α (IRE1α), protein kinase RNA-like ER kinase (PERK), and activating transcription factor 6 (ATF6). They govern together the expression of a vast range of target genes, which are partly overlapping. The proteins encoded by the target genes modulate the adaptation to stress or induce apoptosis (Hetz, 2012).

IRE1α dimerizes and autophosphorylates, which trigger its RNase activity in order to splice the RNA encoding transcription factor XBP1 into its active spliced form XBP1(S). This branch of UPR controls via the active form of XBP1 the genes encoding proteins involved in protein folding and quality control, ER-associated degradation (ERAD), and phospholipid synthesis (Hetz, 2012).

The second branch of UPR, triggered by the activation of PERK, is involved in the attenuation of protein synthesis. The initiation factor eukaryotic translation initiator factor 2α (eIF2α) is phosphorylated by PERK, and the phosphorylation allows the translation of ATF4 mRNA encoding a transcription factor which controls the transcription of genes involved in apoptosis, autophagy, antioxidant responses, and amino acid metabolism (Hetz, 2012).

The third branch of UPR is activated by ATF6, which is transported to the Golgi apparatus, where proteases site 1 protease (S1P) and site 2 protease (S2P) release its cytosolic domain fragment, ATF6f. This fragment contols XBP1 and the upregulation of genes encoding components of ERAD (Hetz, 2012).

2.2.4 Preparation for protein production

Antibodies and other secretory proteins undergo post-translational modifications in the endoplasmic reticulum (ER) to attain their three-dimensional structure. These modifications including glycosylation, disulphide bond formation, and cleavage of the signal sequences, are catalyzed by various chaperones and enzymes (Masciarelli and Sitia, 2008). The differentiation process remodels B cells into antibody-producing 'cell factories' that are built, equipped, and managed for optimal antibody synthesis and secretion. During differentiation, the ER expands into an elaborate network extending thoughout the cytoplasm. The expansion of the ER is complemented by the enlargement of the Golgi complex. The changes needed for converting a B cell into a plasma cell are illustrated in figure 3 (Brewer and Hendershot, 2005). Massive antibody secretion imposes metabolic requirements for the cell, particularly regarding amino acid uptake, ATP synthesis, and redox homeostasis (Masciarelli and Sitia, 2008). Thus the secretory organelles, particularly the ER, are larger in plasma cells than in B cells (Romijn et al., 2005).

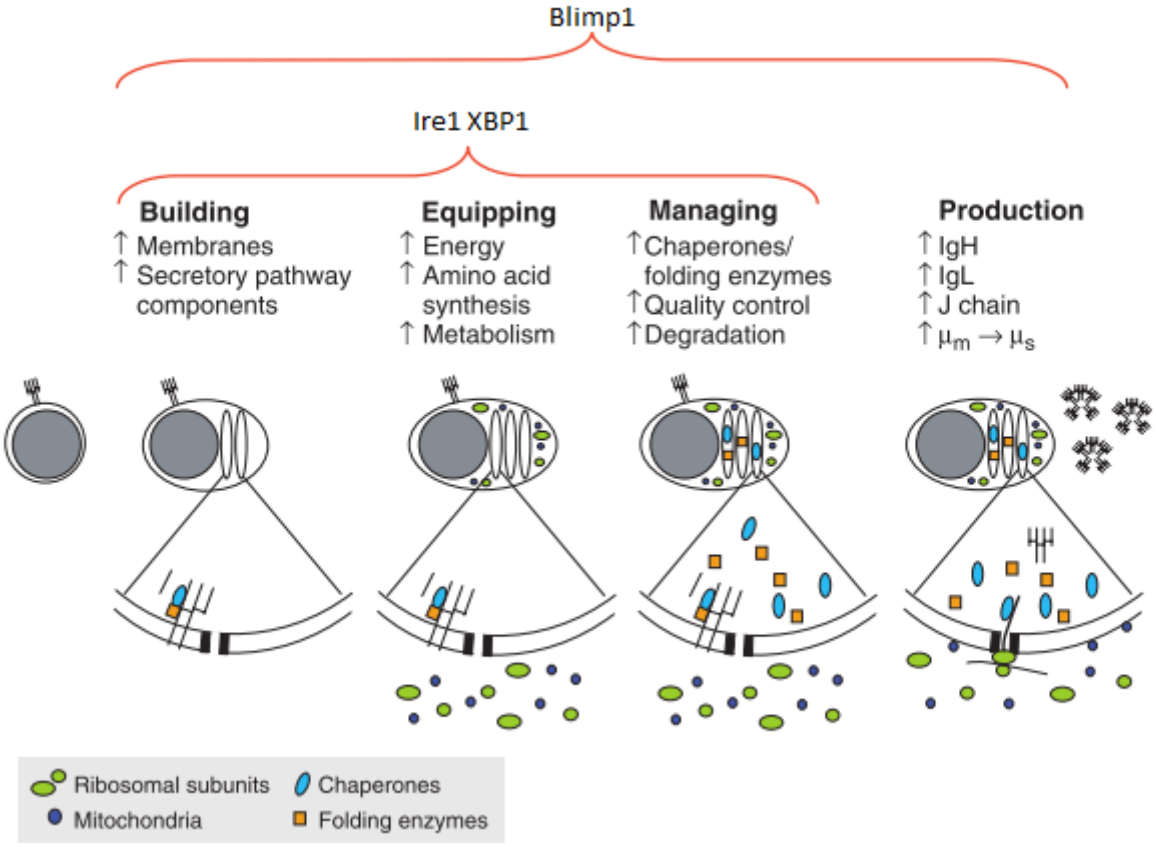

Figure 3: Building an antibody factory requires dramatic remodelling of the cell. The changes include the building of the secretory apparatus, the equipment of the cell with larger amounts of ribosomal subunits, mitochondria to provide energy, and components of metabolic processes and amino acid synthesis, the management of the product quality by increasing the amount of folding enzymes, chaperones, and components of the quality control system, and finally the upregulation of immunoglobulin genes. Blimp1 regulates all of the processes, including the upregulation of XBP1. XBP1 induces proteins involved in the preparation for antibody secretion. Ire1 is needed for remodelling XBP1 into its spliced form. Adapted from Brewer and Hendershot (2005).

Romijn et al. (2005) studied the expression patterns of proteins during plasma cell development. They found that ER-resident proteins were continuously upregulated during the differentiation from B cell to plasma cell. Soon after the activation of naïve B cell, proteins needed for protein production and energy supply were upregulated. Many ER-resident proteins involved in protein folding were upregulated. For instance, protein disulphide isomerase (PDI) was constantly upregulated, and binding immunoglobulin protein (BiP, a molecular chaperone promoting proper protein maturation) showed a transient peak in expression at the first day after B cell activation. Ribosomal subunits, and other proteins involved in translation, were initially strongly upregulated. Proteins involved in the immune system, like Ig light

and heavy chain proteins, were initially downregulated, but upregulated again at later stages of plasma cell differentiation. The production of IgM started when the cell's secretory capacity has increased. These findings show that the B cells have to first prepare themselves for the large scale production of antibodies (Romijn et al., 2005). Plasma cells are able to secrete antibodies at the rate of up to 10^8 IgG, IgA, or IgM molecules per cell per hour (Hibi and Dosch, 1986).

2.2.5 Folding of antibodies

The folding of the heavy and light chains of antibodies begin before the translation of the polypeptide chain is completed. In most cases, the heavy chain dimers are first assembled, and the light chains are added to the dimers covalently via a disulphide bond (Feige et al., 2010). Feige et al. (2010) have illustrated the folding and assembly of an IgG molecule. This is shown in picture 4. The steps are probably true also for other classes of Ig molecules.

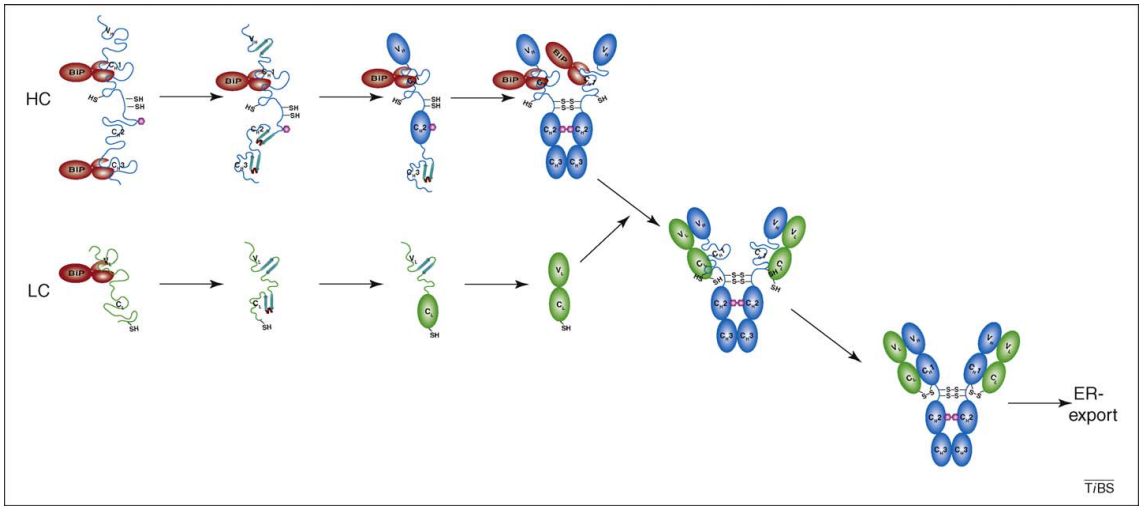

Figure 4: The folding and assembly of an IgG molecule. Folding, glycosylation, and the formation of disulphide bridges starts cotranslationally in the ER. Most of the IgG domains interact with BiP (shown in red) during folding. Most variable domains and all constant domains besides C_H1 fold autonomously. The completion of folding C_H3 induces heavy chain (HC) dimerization solidified by disulphide bridges residing in the hinge region. C_H1 domain remains bound to BiP unfolded until the completed light chain (LC) replaces BiP and C_L induces its folding. Disulphide bridge between the light and heavy chains forms once the C_H1 domain is folded, and the IgG molecule is ready to be secreted (Feige et al., 2010).

In their review, Feige et al. (2010) have grouped Ig domains into three categories on the basis of their folding. In the first category, domains are able to fold to a monomeric state autonomously. The folding is guided by an internal disulphide bridge. A major intermediate on the pathway has a proline residue between two strands that needs to isomerize from its non-native *trans* state to *cis* state before

folding is able to proceed. C_L and C_H2 domains fold this way. The second category, which holds for C_H3 domain of IgG, folds slower than the first category domains and forms an obligate homodimer. It folds via two intermediates, where the second one is formed after a critical proline residue in the first one isomerizes to its native *cis* state. The second intermediate is able to dimerize and complete folding. The third category is the template-assisted folding of the C_H1 domain. The unfolded domain interacts with the C_L domain in the intact antibody, which induces its folding. The prerequisite for folding is not an intramolecular disulphide bridge, like in the other two categories, but a covalent linkage of two cysteine residues (Feige et al., 2010). The same mechanisms which allow the generation of diverse antibodies and the affinity maturation of the immune system also increase the possibility of secreting an antibody which isn't able to fold or assemble correctly. This could affect the functioning of the immune system. Therefore, plasma cells and other B cell lineage cells depend heavily of the quality-control system of the ER in order to ensure the transportation to the cell surface of only correctly assembled antibody molecules (Feige et al., 2010).

3 *Saccharomyces cerevisiae* in antibody production

Yeasts are unicellular eukaryotes that are one of the simplest model organisms. Their sub-cellular organization containing a nucleus, mitochondria, endoplasmic reticulum (ER), Golgi apparatus, secretory vesicles, vacuoles, and microbodies is similar to higher eukaryotes. *Saccharomyces cerevisiae*, or baker's yeast, has been used in human diet for centuries and has been accepted as generally recognized as safe (GRAS). Therefore from the regulatory point of view, it is a good host for producing recombinant protein products. Its biochemical and molecular properties and manipulation techniques are also well known (Berlec and Strukelj, 2013).

S. cerevisiae has beneficial characteristics for protein production of both prokaryotes and eukaryotes. Like prokaryotes, it is able to grow fast in cheap media, and its genetic manipulation is simple. It has the abilities for folding, disulphide bond formation, proteolytic processing, and post-translational modifications without carboxylation, common with eukaryotes. However, *S. cerevisiae* can't reach high cell densities, and its secretory capacity is limited (Berlec and Strukelj, 2013).

3.1 Engineering of *Saccharomyces cerevisiae*

The secretory pathway is a common pathway used to complete secreted proteins and proteins targeted to the plasma membrane and organelles of the endosome membrane system. Secreted proteins are synthesized on ER-bound ribosomes. They are co-translationally translocated into the ER lumen, where they attain their native conformation. After ER, they are transported to the Golgi apparatus by vesicular transport (Anelli and Sitia, 2008; Idiris et al., 2010). In the Golgi apparatus, the proteins are further modified, for instance by glycosylation. Transport vesicles carry the proteins to their destination (Berg et al., 2007).

The production of antibody fragments consisting of the heavy chain variable domain, V_{HH} fragments, in *S. cerevisiae* has been obtained by Frenken et al. (2000) and Thomassen et al. (2002). Frenken et al. (2000) raised in *Lama glama* V_{HH} fragments against human chorionic gonadotropin (hCG), and azo-dyes RR6 and RR120. The highest production level, 9.3 $mg/(l * OD600)$ was reached for one of the hCG antibody fragments, whereas the highest production rate of antibody fragments against RR6 azo-dye was 3.9 $mg/(l * OD600)$. It was found that the secretion rate varied from 25% to 90% for different antibody fragments. Thomassen et al. (2002) were able to produce V_{HH} fragments against the RR6 azo-dye in *S. cerevisiae*. The efficiency of the secretion correlated with the hydrophobicity of the fragments in a way that the more hydrophilic fragments were secreted more efficiently. In total 1.3 kg of V_{HH}s were produced in a $15m^3$ fed-batch fermentation. The obtained specific production rate during induction was $0.213g/(kg * dryweight)$ per hour.

The process of protein folding and subsequent secretion involves many interacting participants. Due to the complexity of the process, modifying the expression level of one step may limit the rate of another one making it a new bottleneck of the whole expression system. Possible bottlenecks limiting the protein yield are shown in figure 5 (Gasser et al., 2008). Factors affecting heterologous protein synthesis

include co- or post-translational translocation of nascent proteins into the ER, protein folding and quality control inside the ER, and protein N-glycosylation in the ER and Golgi apparatus. Protein secretion is also affected by intracellular protein sorting and trafficking, proteolytic degradation, and stress response for misfolding or overexpression (Idiris et al., 2010). Folding, glycosylation, disulfide bond formation, and vesicle trafficking of the secreted protein must all be accomplished, and at the same time quality control feedback loops must be maintained and cellular homeostasis should not be disturbed. Each process has to be tuned to a certain state on the basis of the physical properties of the secreted protein, like size, amount of disulfide bonds, or hydrophibicity. Because of these demands and the complexity of the cellular system, it is difficult to engineer a *S. cerevisiae* strain that could be used for the production of many different recombinant proteins (Hou et al., 2012).

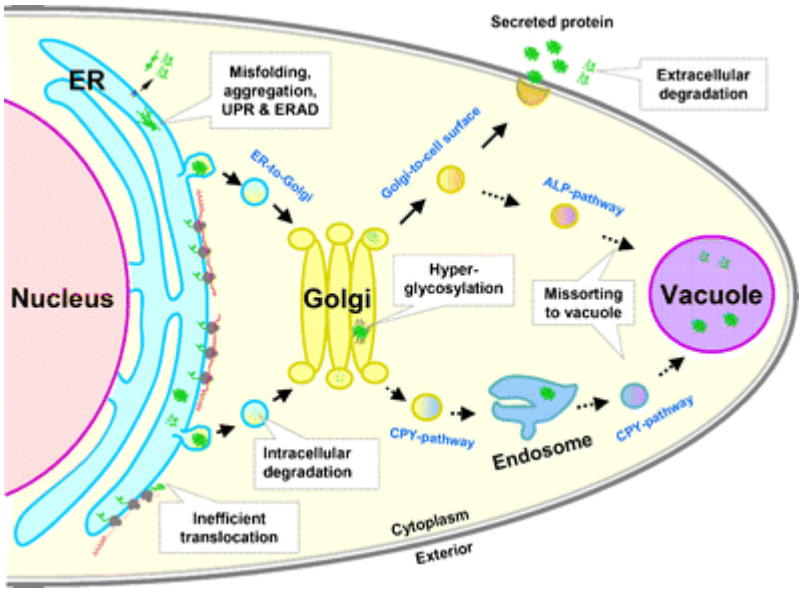

Figure 5: The secretory pathway of *S. cerevisiae* and its potential bottlenecks regarding protein secretion (Idiris et al., 2010).

Current strategies for strain engineering for protein secretion are mainly focused on four topics (Idiris et al., 2010):

- Engineering of the intracellular protein trafficking pathway, including translocation
- Engineering of protein folding and quality control system in the ER
- Engineering of protein N-glycosylation
- Minimization of post-secretory proteolytic degradation

The research focusing on these four topics is reviewed in this chapter.

3.1.1 Engineering of translocation

The intracellular trafficking of secretory proteins is initiated with their translocation into the ER lumen and completed by stepwise vesicular transport. The transport steps after ER are from ER to Golgi, intra-Golgi, and post-Golgi traffic, respectively. After the protein is correctly folded in the ER lumen, the trafficking between organelles is predominantly done by membrane-enclosed transport vesicles. The newly synthesized proteins are selected and concentrated into distinct vesicle populations, which are targeted to a specific acceptor compartment. Inefficient traffic or missorting often results in intracellular retention or accumulation of the target protein for secretion (Idiris et al., 2010). Different proteins can accumulate in different compartments (Hou et al., 2012). When optimizing the intracellular vesicle trafficking pathway, it is important to identify which step is the rate-limiting one (Idiris et al., 2010).

Translocation into ER is determined by the presignal sequence. The presence of the sequence marks the nascent polypeptide chain as one for crossing the ER membrane, and its hydrophobicity determines whether the polypeptide chain is directed to the ER lumen by the signal-recognition particle (SRP) (Ng et al., 1996). Thus engineering of the signal sequence is a way to affect the translocation. In the SRP-dependent route, the signal sequence is recognized by the signal-recognition particle (SRP), which binds to the sequence and the ribosome during translation, which is then paused. SRP binds to the SRP receptor (SR) on the ER membrane, and the SRP-SR complex brings the ribosome to the ER membrane, where it docks with a protein-conducting channel, the translocon. Protein synthesis resumes, and the growing polypeptide chain passes through Sec61 complex of the translocon into the ER lumen (Berg et al., 2007). If SRP doesn't bind strongly enough to the signal sequence on the polypeptide chain, translation keeps proceeding, and the nascent polypeptide chain is targeted posttranslationally to the translocon through Sec62/63p complex. Then, like the polypeptides interacting with SRP, the chain is passed through Sec61 complex into ER lumen (Ng et al., 1996).

Synthetic leader sequences can be used to overcome problems arising from inefficient processing of pre- or pro-leaders, hyperglycosylation, protein accumulation, or incorrect trafficking in the secretory pathway. Shusta et al. (1998) used two different synthetic leader sequences in their study for increasing scFv production in *S. cerevisiae*. The first one was a synthetic pre pro region based on a consensus signal sequence, and the construct was expressed under the control of GAL 1-10 promoter. The second one was a synthetic signal peptide derived from α factor, expressed under under the control of constitutive GAPDH promoter. The first one resulted in approximately 6- to 10-fold increases in protein production over low copy plasmid levels, but after overexpression of Gal4p, which increases transcription from the GAL 1-10 promoter, the secretion levels dropped due to the oversaturation of the secretory pathway because of scFv accumulating in an ER-processed pre-Golgi form. The second one resulted in elevated secretion levels, as described in section 3.1.2 (Shusta et al., 1998).

3.1.2 Engineering of protein folding in the ER

Possible bottlenecks in the ER are protein misfolding, protein aggregation, unfolded protein response, and ER-associated protein degradation (see figure 5) (Idiris et al., 2010). The inability of foreign polypeptides to reach their native conformation in heterologous host cells usually leads into their prevalence in the insoluble cell fraction. The cells are driven to a global conformational stress condition by the abnormally high and non-physiological rates of recombinant protein production and the abundance of misfolded protein species. By a series of individual physiological responses the cell aims to restore cellular homeostasis and minimize any toxicity of misfolded protein species (Gasser et al., 2008). The accumulation of unfolded or misfolded protein triggers the unfolded protein response (UPR) pathway, which aims at reducing the ER stress by inducing genes involved in protein folding and ER-associated protein degradation (ERAD) (Delic et al., 2014a). Idiris et al. (2010) have reviewed ways to overcome bottlenecks associated with ER, and they have concluded that overexpressing the genes encoding protein disulfide isomerase (Pdip) and chaperones like Kar2p, which is the yeast homologue of binding immunoglobulin protein (BiP) (Rose et al., 1989), which binds to the nascent polypeptide, can be an effective strategy to increase protein secretion rates in *S. cerevisiae*.

A study on the effects of Pdip and Kar2p on protein secretion in *S. cerevisiae* was done by Xu et al. (2005). The purpose of the study was to examine unfolded protein response during single-chain antibody fragment (scFv) 4-4-20 expression in *S. cerevisiae*. It was found that the overexpression of *PDI* reduced UPR significantly, and increased the secretion of scFv. Co-overexpression of both *PDI* and *KAR2* resulted in a significant increase in the secreted levels of scFv. The proposed reason for the improved expression of scFv was that the overexpression of *KAR2* led to improved translocation of the newly synthesized polypeptide, because it binds to scFv during its translocation, and the overexpression of *PDI* resulted in improved folding rate of scFv. The overexpression of only *KAR2* resulted only in modest increase of scFv secretion, and it was not found to reduce the UPR (Xu et al., 2005). A similar study was done by Shusta et al. (1998), who studied the effects of co-overexpression of *KAR2* and *PDI* on production of single-chain antibody fragments in *S. cerevisiae*. The co-overexpression resulted in 2-8 fold increase of secretion titers for the studied five scFvs.

Hackel et al. (2006) produced anti-transferrin receptor single-chain antibody (OX26 scFv) in *S. cerevisiae*. They modulated the expression temperature and overexpressed chaperones in the ER in order to optimize the production yield. The optimal temperature was found to be +20 °C. Of the chaperones in the ER, the best results were achieved when co-overexpressing both *KAR2* and *PDI*. The achieved purified yield of OX26 ScFv was 0.5 mg/l.

Recently, de Ruijter and Frey (2015) researched the effects of disrupting ER protein quality control on the antibody production in *S. cerevisiae*. They deleted individually *YOS9, HTM1, UBC7, HRD1, or HRD3* genes. The deletions were done both with and without disrupting the UPR by deleting *IRE1* gene. Antibody production was slightly increased with the *HTM1* deletion. Other deletion strains

showed decrease in the antibody production, especially the Δ*yos9* strain.

Modifying one step in the secretory pathway may lead to rate limitation of the following one, which may then become the bottleneck of the secretory pathway. Additionally, the effects of modification are usually protein specific. Thus, the secretion of different heterologous proteins may be improved by overexpressing multiple folding helpers and chaperones, or by targeting heterologous or cytosolic chaperone to the ER lumen (Idiris et al., 2010).

3.1.3 Engineering of protein glycosylation

Protein glycosylation is a significant modification process in the secretory pathway of *S. cerevisiae* (Idiris et al., 2010). Initial glycosylation happens already during translocation. Glycosylation aids folding of the protein, it protects the protein from proteases, and works as a signal for quality control (Hou et al., 2012). Native full length monoclonal antibodies are glycosylated during their synthesis, and the glycosylated constant region stabilizes the antibody and is important for its biological functioning. Glycosylation of antibodies also impacts their clearance rate from the body, and incompatibly glycosylated molecules may cause severe immunogenic effects (Spadiut et al., 2014). Due to the different *N*-glycan structure, antibodies produced in *S. cerevisiae* may cause immunogenic reactions and be cleared rapidly from the bloodstream after interacting with human mannose receptors (Young and Robinson, 2014).

In yeast, there are two types of glycosylation, *N*-linked and *O*-linked. The glycans involved, *N*-glycan ad *O*-glycan are pictured in figure 6. In *N*-linked glycosylation, a 14 sugar glycan tree is added to asparagine residue of the recognition sequence (N-X-S or N-X-T, X is any amino acid except proline). The anchor of the glycan tree, which is attached to the asparagine residue, is a *N*-acetylglucosamine. The ER-resident oligosaccharyl transferase (OST) completes the glycosylation. *O*-linked glycosylation occurs at the hydroxyl groups of threonine and serine, and it is catalyzed by protein *O*-mannosyltransferases (PMTs). A single mannose is transferred to the serine/threonine in the ER by PMTs (Hou et al., 2012). These steps happen in the ER, and they are similar in yeasts and humans. The differences arise from the glycosyltransferase reactions in the Golgi apparatus. In yeast, the *N*-glycan intermediate is modified by several mannosyltransferases that attach more than 50 mannose residues to it (Chiba and Akeboshi, 2009).

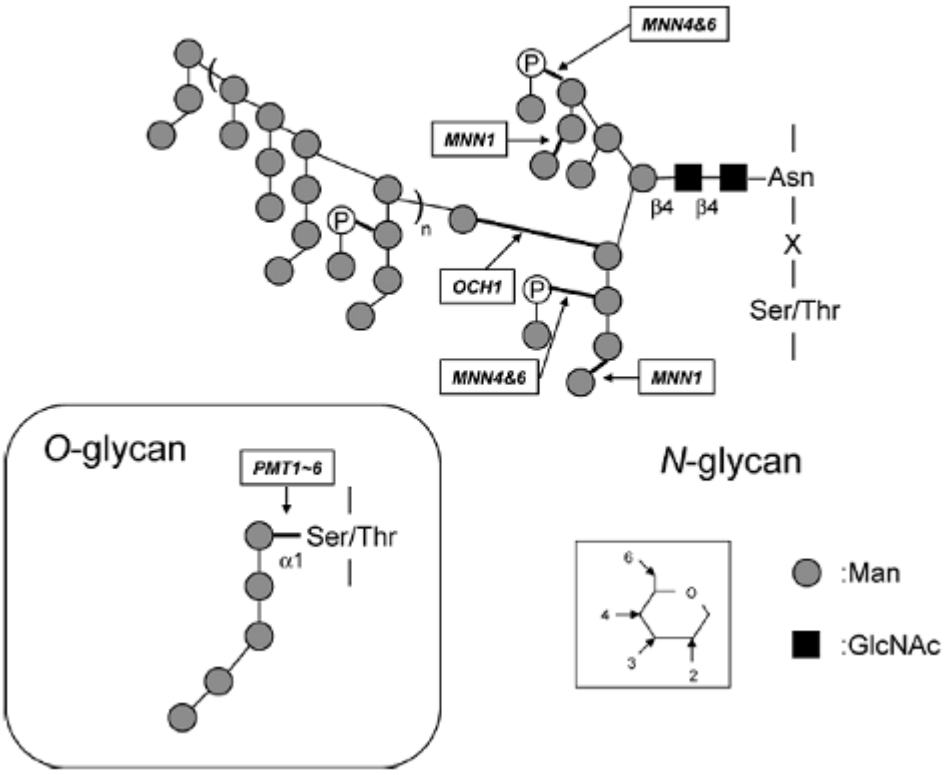

Figure 6: *N*- and *O*-glycan structures of *S. cerevisiae*. Beta linkages are indicated in the figure, all other linkages are alpha-anomeric. Each linkage is indicated as a number in the frame on the bottom right. The genes which encode main enzymes related to yeast-specific modification are represented in bold (Chiba and Akeboshi, 2009).

The differences of *N*-type glycosylation in *S. cerevisiae* and mammals are illustrated in figure 7, along with a pathway resulting in a human-like *N*-glycan in *S. cerevisiae* engineered by Parsaie Nasab et al. (2013).

A novel synthetic pathway for producing $GlcNAc_2Man_3GlcNAc_2$, a human-like *N*-glycan structure, in *S. cerevisiae* was recently engineered successfully by Parsaie Nasab et al. (2013). The structure was confirmed to be present on the secreted antibody, which was a hybrid human IgG monoclonal antibody directed against hen egg lysozyme (HyHEL). The strain used was a Δ*alg3* Δ*alg11* double mutant, where the *N*-glycans are not matured to their high-mannose native structure. The deletions abrogate lipid-linked oligosaccharide (LLO) synthesis, and give rise to lipid-linked $Man_3GlcNAc_2$ structure, which is built up on the cytoplasmic side of the ER in the mutant strain and flipped by an artificial flippase into the ER lumen, where it is transferred to the nascent polypeptide by a protozoan oligosaccharyltransferase. Protein-bound $Man_3GlcNAc_2$ is a substrate for human N-acetylglucosaminyltransferases I and II (GnTI and GnTII) targeted in the Golgi, giving rise to complex *N*-glycan $GlcNAc_2Man_3GlcNAc_2$. The pathway is illustrated in figure 7, along with the *N*-glycan pathways in mammals and yeast (Parsaie Nasab et al., 2013).

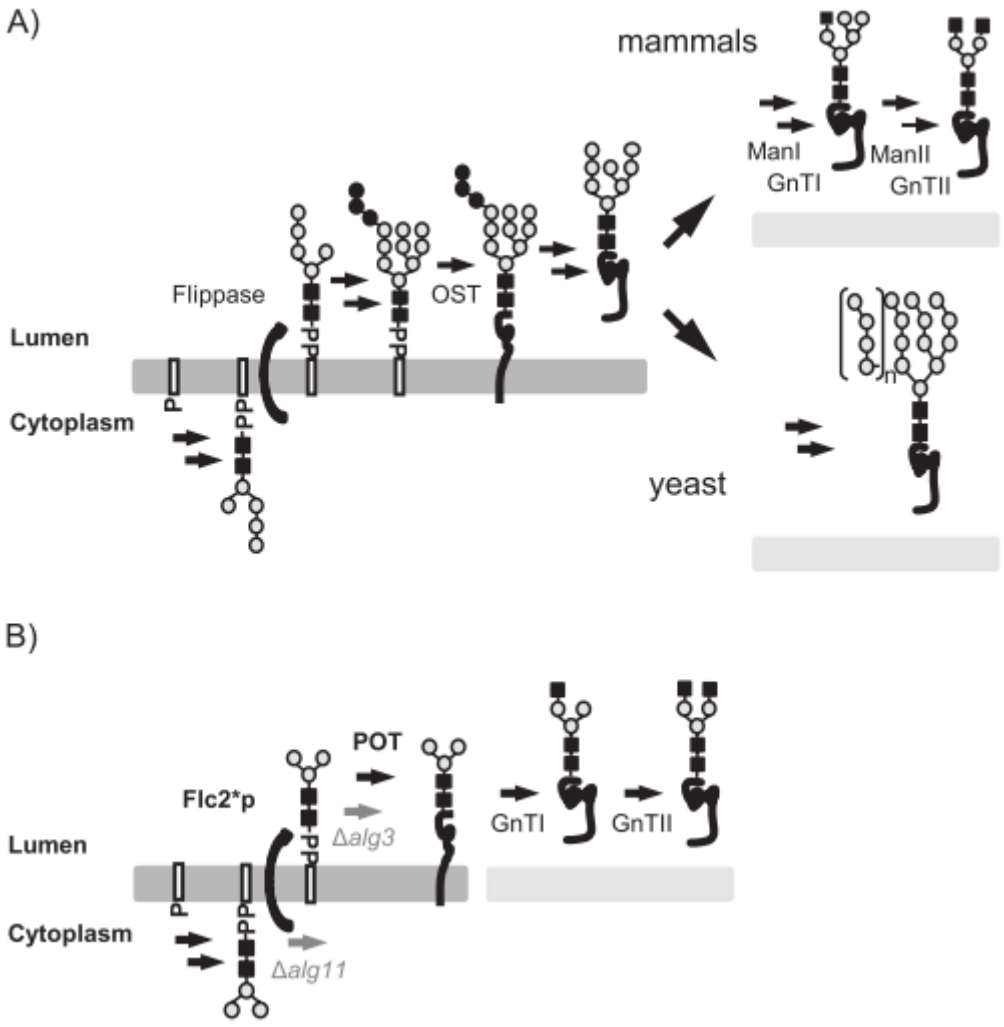

Figure 7: The comparison of *N*-glycan biosynthesis pathways in mammals and yeast, and of the artificial *N*-glycan biosynthesis pathway in *S. cerevisiae* engineered by Parsaie Nasab et al. (2013). (A) Lipid-linked $Man_5GlcNAc_2$ is flipped into the ER lumen from the cytoplasmic face of the ER membrane. In the ER, it is further elaborated by luminal mannosyl- and glucosyltransferases. The resulting structure, $Glc_3Man_9GlcNAc_2$, is transferred to the nascent polypeptide chain by oligosaccharyltransferase (OST), and three glucose and one mannose residue are trimmed. In mammals, mannosidase I (ManI) trims three mannose residues in the Golgi apparatus, resulting in a substrate for GnTI. A substrate for GnTII is generated by mannosidase II (ManII), which trims two additional mannose residues. In *S. cerevisiae*, a number of mannosyltransferases catalyze the attachment of mannosyl residues, resulting in highly mannosylated *N*-glycans. (B) The engineered pathway as described in the text above. Flc2*p is the artificial flippase, and POT is protozoan oligosaccharyltransferase. Mannose residues don't need to be trimmed in this pathway. Grey circles are mannoses, black circles are glucoses and black squares are *N*-acetylglucosamines. The dark grey bar is the ER and light grey bar is the Golgi apparatus. Adapted from Parsaie Nasab et al. (2013).

As reviewed by Piirainen et al. (2014), the manipulation of LLO biosynthesis results often in the accumulation of intermediate structures and hypoglycosylation of target proteins. The efficiency of *N*-glycosylation can be improved by the combined use of protozoan oligosaccharyltransferases and the overexpression of artificial flippase, which were both done by Parsaie Nasab et al. (2013).

The first reported recombinant *S. cerevisiae* able to produce an intermediate *N*-glycan structure identical with the human structure was engineered by Chiba et al. (1998). This $Man_5GlcNAc_2$, which is a human-like intermediate for hybrid- and complex-type sugar chains, was engineered by transforming a triple mutant *S. cerevisiae* strain lacking three mannosyltransferase activities with an α-1,2-mannosidase (from *Aspergillus saitoi*) expression vector (Chiba et al., 1998). The disrupted genes were *OCH1*, *MNN1*, and *MNN4*,encoding α-1,6-mannosyltransferase, α-1,3-mannosyltransferase, and a positive regulator for phophomannosyltransferase (Mnn6p), respectively (Chiba et al., 1998; Nakanishi-shindo et al., 1993).

In contrast to yeast, there are several types of *O*-glycosylation in mammals. These include for instance mucin-type (*O*-linked *β*-*N*-acetylgalactosamine: *O*-GalNAc), *O*-linked *β*-*N*-acetylglucosamine (*O*-GlcNAc), *O*-linked fucose (*O*-Fuc) and *O*-linked glucose (*O*-Glc). *O*-GalNAc modifications occur on serine and threonine side chains of nuclear and cytoplasmic proteins that are involved in almost all cellular functions (Chiba and Akeboshi, 2009). Mucins and mucin-like glycoproteins are also potential novel cancer markers, and thus the production of mucin-like glycoproteins for induction of specific antibodies could be useful (Amano et al., 2008). *O*-Fuc and *O*-Glc are found in several serum proteins. They are post-translational modifications that are relevant in the early stages of development and are vital for certain proteins' physiological functions (Chiba and Akeboshi, 2009).

Amano et al. (2008) engineered a strain of *S. cerevisiae* capable of producing mucin-type sugar chains. This was done by transforming genes encoding human UDP-Gal/GalNAc transporter, human ppGalNAc-T1, *Bacillus subtilis* UDP-Gal/GalNAc 4-epimerase, and *Drosophila melanogaster* core1 *β*1-3 GalT into *S. cerevisiae*. The resulting yeast was capable of producing a MUC1 like peptide containing *O*-glycan, and a mucin-like glycoprotein human podoplanin, which is a platelet-aggregating factor. The pathway is illustrated in figure 8 (Amano et al., 2008).

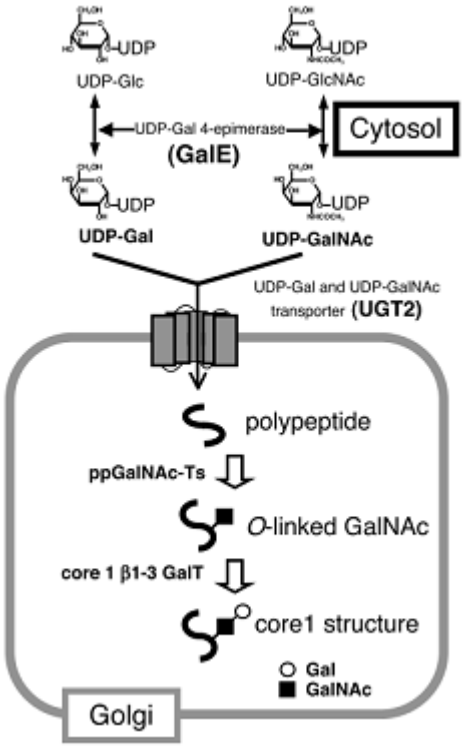

Figure 8: A mucin-type pathway in *S. cerevisiae*. In the cytosol, UDP-Gal and UDP-GalNac are synthesized from UDP-Glc and UDP-GlcNAc, respectively, by GalE protein. UDP-Gal and UDP-GalNAc are transported from the cytosol to the Golgi lumen by UTG2 protein. In the Golgi, ppGalNAc-Ts and core1 β1-3 GalT transfer GalNAc and Gal to polypeptides (Amano et al., 2008).

Chigira et al. (2008) used an artificial *O*-glycosylation pathway in *S. cerevisiae* for producing an *O*-fucosylated epidermal growth factor (EGF) in yeast. The *O*-fucosylation system provides a way to produce proteins with homogenous carbohydrate chains. These proteins can be used for the production of antibodies recognizing *O*-glycosylated EGF domains. The system was engineered via expression of genes that encode the EGF domain and protein *O*-fucosyltransferase 1, along with genes which protein products convert cytoplasmic GDP-mannose to GDP-glucose (Chigira et al., 2008).

3.1.4 Minimization of post-secretory proteolytic degradation

Host-specific proteases are present in relatively high levels in *S. cerevisiae*. The post-secretory degradation of recombinant gene products caused by them is one of the major problems hindering the effective secretion and purification of heterologous proteins. The proteases are readily induced by environmental stresses, especially during high-density fermentation processes. The effects of proteases can be lessened by either changing the cultivation conditions like temperature or pH, changing medium composition, adding protease inhibitors, or by genetic manipulation of the

host proteases. The first three approaches are usually protein specific, and the last one can be used to develop protease-deficient strains of *S. cerevisiae* (Idiris et al., 2010).

An example of successfully manipulating the cultivation conditions was described in a study by Kang et al. (2000). They researched the secretion of recombinant human serum albumin (HSA) from the GAL10 promoter in *S. cerevisiae*. The target protein underwent rapid degradation, which was accelerated by carbon-source feeding. It was noted that the degradation correlated closely with the acidification of extracellular pH, and therefore it was overcome either by buffering the culture medium above pH 5.0, or by preventing the acidification of medium pH by adding amino-acid rich supplements to the culture medium. It was concluded that HSA is very susceptible to pH-dependent proteolysis, which is mediated by cell-bound proteases whose activity and expression can be affected by modifying the cultivation medium (Kang et al., 2000).

A study by Jønson et al. (2004) was an example of minimizing proteolytic degradation by gene distruption. They identified a novel endoprotease in *S. cerevisiae*, Cym1p, which is a member of the pitrilysin family. The target peptide tested was pro-cholecystokinin (proCCK), which is proteolytically processed by Cym1p. Disrupting the gene coding Cymp1, *CYM1*, reduced intracellular proteolysis and increased the secretion of proCCK. It was also found that the disruption increased the secretion of growth hormone and pro-B-type natriuretic peptide (Jønson et al., 2004).

4 Transcriptomics and gene enrichment analysis

Transcriptomics is one of the 'omic' technologies that aim to adopt a holistic view of the molecules in a cell, tissue, or organism (Horgan and Kenny, 2011). It studies the transcriptome, which is the total set of RNA transcripts the genome produces under certain circumstances in a certain cell (Nature Publishing Group, 2014). According to Wang et al. (2009), the main aims of transcriptomics are:

- Cataloguing all transcript products, mainly mRNAs, but also small RNAs and non-coding RNAs
- Determining the transcriptional structure of genes including their start sites, 5' and 3' ends, splicing patterns, and post-transcriptional modifications
- Quantifying the changing expression levels of each transcript during development and under different conditions

Transcriptomics makes possible the identification of differentially expressed genes between different samples, and the significance of those genes can be further examined by gene enrichment analysis (Huang et al., 2009; Nature Publishing Group, 2014).

4.1 Microarray and RNA-Seq experiments

A microarray analysis is a way to find out the expression levels of even tens of thousands of genes in a single assay (Quackenbush, 2006). It has made the rapid and high-throughput quantification of the transcriptome possible. Results from microarray experiments have provided valuable information about the transcriptome in different cell types and the changing levels of gene expression under different conditions (Malone and Oliver, 2011).

A microarray experiment is conducted by arraying seuqnce-specific probes, which represent thousands of individual genes, on an inert substrate (Quackenbush, 2006). The design of the probes is usually based on genome sequence or on open reading frames. RNA is extracted from cells of interest and labelled with one or two colours of fluorescent dye, or other detectable marker. (Malone and Oliver, 2011) Each RNA hybridizes with its complementary probe, and the relative fluorescence intensity of each gene-specific probe is measured to indicate the concentration of the RNA. In order to be able to do comparisons between assays, the obtained data is normalized in order to compensate for differences in hybridization, labelling, and detection efficiencies. After normalization, the data can be filtered in a number of ways, for instance genes that have minimal variance across the samples can be left out from further inspection. The collected, normalized, and filtered data is then ready for analysis like gene enrichment analysis (Quackenbush, 2006).

RNA-Seq is a new method of analysing the transcriptome by a procedure known as deep sequencing. Transcripts present in the starting material are sampled by direct sequencing, instead of using the hybridization technique of microarrays. The sequences are mapped back to a reference genome, and the number of mapped reads is the measure of the expression level for a gene (Malone and Oliver, 2011).

Compared to microarrays, RNA-Seq has several advantages. Junctions between exons can be assayed without prior knowledge of the gene structure, since RNA-Seq provides direct access to the sequence. Also RNA editing events can be detected. RNA-Seq can also be used on species for which a full genome sequence is unavailable. Expressed regions of the genome corresponding to currently unidentified genes can be easier to detect with sequencing than with microarrays. RNA-Seq makes also quantifying individual transcript isoforms easier when compared to microarrays (Malone and Oliver, 2011).

The pros of microarray experiments compared to RNA-Seq are the quality of the data and cheap price. As microarray experiments have been conducted for a long time, their biases are understood nowadays. In the early experiments different microarrays seemed to give different results from the same samples. The fluorescent readout of the intensities varied between laser scanners. Recognition of biases has led to the development of quality procedures, for instance experimental and computational methods have been developed for dealing with variations between laboratories. This work is still being done with RNA-Seq. Microarray experiments cost also about one tenth of RNA-Seq experiments. Inadequate coverage of RNA-Seq may be a problem because of the high amount of expensive reads needed for saturated coverage. This is not a problem for microarray experiments due to the fixed nature of probes (Malone and Oliver, 2011).

4.2 Proteomics and transcriptomics

The set of all expressed proteins in a cell, organism, or tissue is called the proteome (Horgan and Kenny, 2011). It has been assumed, that there is correlation between mRNA transcripts and protein expression (Haider and Pal, 2013). However, studies by for instance Gygi et al. (1999) and Ghazalpour et al. (2011) have shown that there can be low correlation between them. Haider and Pal (2013) have summarized different factors affecting the relationship between mRNA and protein expression levels. Factors which influence the efficiency of translation, for instance physical properties of the transcript, will also influence the correlation between mRNA and protein. Codon-bias, the way how a number of codons can be used in translating the same amino acid, also impacts the correlation. Ribosome-density, the number and distribution of ribosomes in a transcriptional unit, strongly influences the efficiency of translation. During the cell cycle, high variability of the expression levels of mRNA results in higher correlation with the protein expression. Eukaryotic mRNA and proteins have different half-lives, which influences the correlation (Haider and Pal, 2013).

There are different methods for integrating proteomic and transcriptomic data. A reference data set can be created from a union of proteomic and transcriptomic data sets. Another way is to extract common functional context of proteomic and transcriptomic data, for instance by utilizing the Gene Ontology. Topological networks may be used to find common regulators from proteomic and transcriptomic data sets. Correlation of transcriptomic and proteomic data can be enhanced also by merging multiple proteomic and transcriptomic data sets respectively, and conducting

a correlation analysis between the merged data sets. Missing values of proteomic data can be predicted by non-linear or linear optimization. Another mathematic model is multiple regression analysis, which can be used to analyse the contribution of sequence features in the correlation between mRNA and protein expression, as the protein abundance depends on other biological factors besides only the abundance of mRNA. Also clustering and dynamic modelling can be used to compare proteomic and transcriptomic data (Haider and Pal, 2013).

4.3 Gene Ontology

An ontology is a vocabulary that represents and communicates knowledge about a topic in addition to a set of relationships between the terms of the vocabulary. Gene Ontology (GO) project, maintained by Gene Ontology Consortium, aims to construct a structured, precisely defined, controlled vocabulary that has consistent descriptions of genes and gene products regardless of species. Since a structured ontology can be easily dealt with by a computer, one of the main uses of GO is performing gene enrichment analyses that find out which GO terms are enriched in a study (Ashburner et al., 2000).

GO terms can be attributes of genes, gene products, or gene-product groups. GO annotations assign GO terms to gene products. The terms are somewhat hierarchical, with 'child' terms being more specific than their 'parent' terms, but unlike in a strict hierarchy, one term can have multiple parent terms (Ashburner et al., 2000).

GO is divided into three categories: biological process (BP), molecular function (MF), and cellular component (CC). Biological process involves often a physical or chemical transformation that is accomplished by molecular functions. It is an objective to which the gene or its product contributes. For instance 'immune response' and 'regulation of B cell proliferation' are BP terms. Molecular function is a biochemical activity of a gene product. For instance 'enzyme binding' and 'MHC class II receptor activity' are MF terms. Cellular component is the part of the cell where a gene product is active. For instance 'endoplasmic reticulum' and 'Golgi apparatus' are cellular component terms. A protein can take part in many processes, contain diverse domains with different molecular functions, and interact with multiple other proteins or locations in the cell, thus there are plenty of relationships between a gene product and GO categories. Figure 9 shows a network around BP term 'immune response-regulating signalling pathway' (Ashburner et al., 2000).

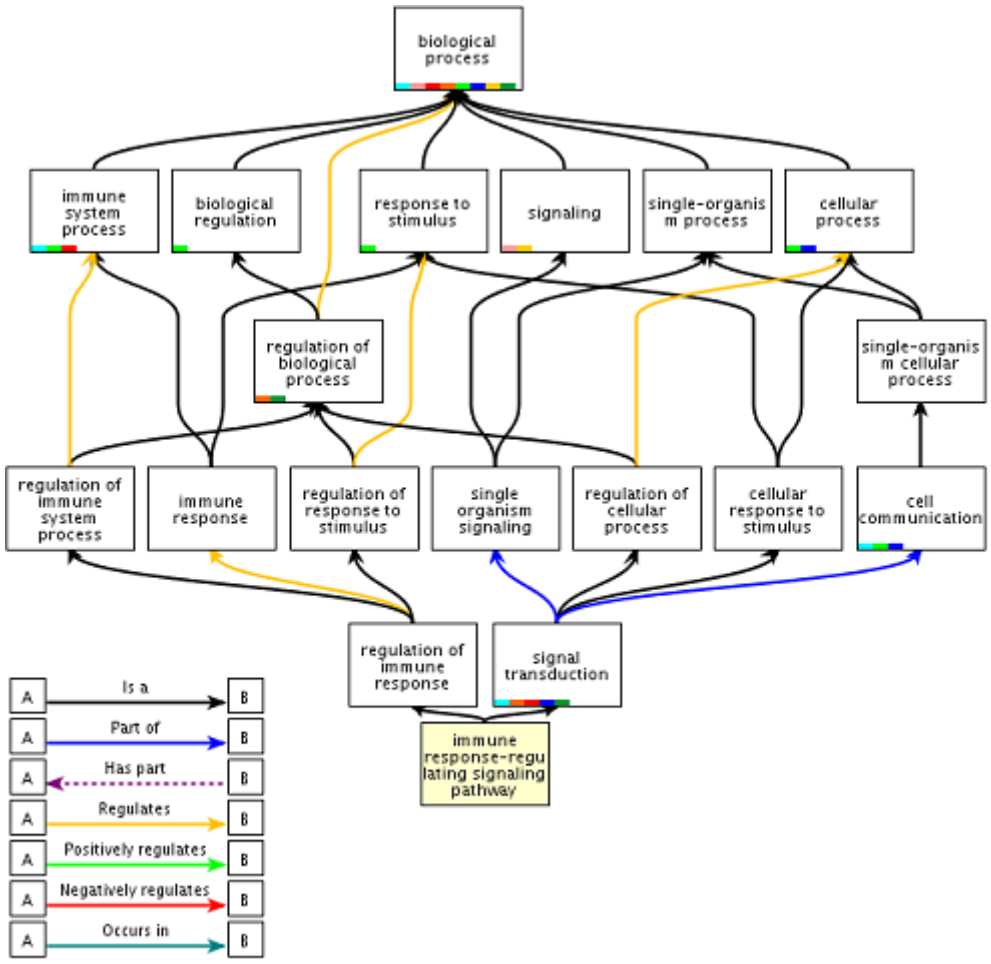

Figure 9: An example of a biological process ontology which illustrates the complex relationships between different terms (Ashburner et al., 2000).

4.4 Gene enrichment analysis

Gene enrichment analysis is a way to find out which biological functions are the most relevant in a given study. The idea is that if a certain process, like immune response, is significant to a given study, its co-functioning genes should be enriched, i.e. be more likely to be selected as a relevant group by statistical methods (Huang et al., 2009). For instance, Cocco et al. (2012) studied different clusters of genes in the generation of long-lived plasma cells, and found that in a certain cluster a significant amount of genes were attributed with B cell activation. They were thus able to conclude that B cell activation was enriched within that cluster of genes.

The enrichment value of a GO term can be quantitatively measured by statistical methods like Hypergeometric distribution, Chi-square, Binomial probability, and Fisher's exact test. It must always be measured against a reference background, which can be the total genes in a genome, or a narrower set of genes like those detected by a microarray. Within the same analysis, the reference background should

be the same, so the results are consistent with each other. There are three ways to conduct an enrichment analysis, singular enrichment analysis (SEA), gene set enrichment analysis (GSEA), and modular enrichment analysis (MEA) (Huang et al., 2009).

A singular enrichment analysis is conducted by first finding out which genes are differentially expressed between experiment and control samples, and then testing the enrichment of each annotation term one-by-one in a linear mode. Differentially expressed genes can be found by t-test, fold change analysis, or their combination. The enrichment P-value describes the probability of finding a certain amount of genes associated with the same GO term compared to random chance. Singular enrichment analysis is able to efficiently extract the major biological meaning behind large gene lists (Huang et al., 2009).

Student's t-test is an approach to find out if a gene is differentially expressed between different conditions. The test is based on the t-statistic assessing the signal-to-noise ratio of the gene in question and comparing its expression measure for the two experimental conditions. T-statistic is defined as 1, where X_i is average value across measurements of the signal-to-noise ratio measured under condition i, and σ_i^2 is the standard deviation of the mean (Quackenbush, 2006).

$$t = \frac{\text{signal}}{\text{noise}} = \frac{\text{difference between groups}}{\text{variability of groups}} = \frac{\langle X_A \rangle - \langle X_B \rangle}{\sqrt{\sigma_A^2 + \sigma_B^2}} \tag{1}$$

There are different ways to use the t-statistic to estimate the probability of a gene being differentially expressed between the conditions. One approach is to use the properties of the t-distribution for normally distributed random variables in order to calculate the probability that for a gene under two conditions (A and B), the two distributions of the expression measures overlap for a given value of t. This is based on the assumption that the variabilities of these measurements follow normal distribution, which is not necessarily always the case for gene expression. However, the t-test is quite robust to violations of the assumption of normality (Quackenbush, 2006).

Another approach is performing a permutation test that uses the properties of the expression measures themselves to estimate the significance of a given t-value. Expression level measurements between the two expression groups are randomly swapped up to the total amount of possible unique permutations. A value for t is calculated each time, and it is determined how often an equal or larger t-value occurs than the one measured for real data. This allows the estimation of the probability that there is a significant separation between the two expression groups. The probability can be expressed as P-value, and a cutoff P-value can be defined to be used for filtering genes for further analysis. For example, if the P-value cutoff is 0.05, then there is at least 95% chance that a gene can have distinguished expression levels between groups (Quackenbush, 2006).

One statistical method for conducting singular enrichment analysis is Fisher's exact test which utilizes the Hypergeometric distribution works in a following way: In a gene universe, all genes can be marked either significant or insignificant to a given study. A null hypothesis is that there is no relationship between a gene being

significant and being in a given GO category containing K genes. Under the null hypothesis, the amount of significant genes can be modelled by Hypergeometric distribution. If there are j significant genes in the given GO category, it is possible to calculate the probability of getting j or more significant genes in K draws from the gene universe, without replacement (Falcon and Gentleman, 2008).

Gene set enrichment analysis is similar to singular gene enrichment analysis, but it takes into account all genes from a microarray experiment. This way all information from microarray experiment is used, and the factors of the gene selection step do not impact the outcome of the analysis. It also evaluates the data at the gene set level, where the sets are formed on the basis of biological knowledge. The goal of the analysis is to find out whether the members of a gene set show differential expression between two biological states. This is done by ranking the genes according to the difference in their expression, and determining where on the list the genes in a gene set fall into. The calculated enrichment score reflects the degree to which the gene set is overrepresented in either the top or bottom part of the ranked list. The significance, or the P-value, of the enrichment score is estimated by a permutation test (Subramanian et al., 2005).

The third way of doing enrichment analysis is modular enrichment analysis, where the basic enrichment calculation is similar to singular enrichment analysis. In addition to singular enrichment analysis, term-to-term relationships are considered by using genes that have joint GO annotation terms as a reference background. These joint terms may have unique biological meaning that is not discovered by looking at only individual GO terms. The downside of the modular enrichment analysis is that terms or genes which have weak relationships to their neighbours may be excluded from of the analysis (Huang et al., 2009).

This P-value given by enrichment analysis can be regarded as an advising scoring system between enriched terms. The P-values themselves can be fragile and influenced by factors such as the used statistical method, the data source, or the studied biological phenomenon itself. Therefore, instead of only looking at P-values, also biological knowledge plays an important role in interpreting the results of enrichment analysis, in a way that the result should make sense based on *a priori* biological knowledge. Also the size of the gene list influences P-values, so that a larger gene list can have higher statistical power, and subsequently higher sensitivity leading to more significant P-values. Thus, it is difficult to compare the absolute P-values of different gene lists (Huang et al., 2009).

5 Scope of the research

The target of this thesis was to find the enriched GO terms in plasma cells, the genes comprising those terms and study the effects of those genes on antibody-secreting *S. cerevisiae*. The enriched terms and the genes comprising them were found by computational enrichment analysis. The selected genes were individually overexpressed and knocked out in antibody-secreting *S. cerevisiae*.

6 Materials and methods

6.1 Bioinformatics

The processing and analyzing of data was conducted in R software environment (R Core Team, 2014). The purpose of the bioinformatics part was to find out which GO-terms were enriched in plasma cell development, which genes comprised those terms, and out of those genes, which ones were applicable for testing in *S. cerevisiae*.

6.1.1 Data selection and pre-processing

Gene expression data from four studies was used in this thesis. All data sets have been deposited in the Gene Expression Omnibus (GEO) database maintained by the National Center for Biotechnology Information (NCBI) (Edgar et al., 2002). The data sets and their respective accession numbers are listed in table 1. All selected data sets contain data from both naïve B cells and matured plasma cells, and these states are compared in the enrichment analysis.

Table 1: Data sets used in this study.

Data set	**Contributors**	**GEO accession number**
Expression profiles from a variety of resting and activated human immune cells	Abbas et al. (2005)	GSE22886
In vitro generation of long-lived human Plasma Cells	Cocco et al. (2012)	GSE41208
Analysis of proposed human B1 cells	Covens et al. (2013)	GSE42724
Gene expression by human splenic B-cell subsets	Good et al. (2009)	GSE13411

Data sets were downloaded to R using the GEOquery package (Davis and Meltzer, 2007) and quantile normalized with the affyPLM package (Bolstad, 2004), except the already normalized data of Cocco et al. (2012). Probes without EntrezID were removed. Data sets from Good et al. (2009) and Abbas et al. (2005) were log2-transformed, and the other two sets had already been transformed. Nine data subsets were formed from the original data sets based either on the age of the plasma cells, or whether they were harvested from peripheral blood, human spleen, or bone marrow. The subsets and their abbreviations are described in table 2.

Table 2: Data subsets and their abbreviations used in this thesis.

Data subset	**Abbreviation**
Abbas et al. (2005) subset peripheral blood A	A1
Abbas et al. (2005) subset bone marrow A	A2
Abbas et al. (2005) subset peripheral blood B	A3
Abbas et al. (2005). subset bone marrow B	A4
Cocco et al. (2012), subset peripheral blood, day 20	CC1
Cocco et al. (2012), subset peripheral blood subset day 41	CC2
Cocco et al. (2012), subset bone marrow	CC3
Covens et al. (2013). subset peripheral blood	C1
Good et al. (2009), subset human spleen	GT1

Annotations for the Illumina HumanHT-12 V4.0 expression beadchip, Affymetrix Human Gene 1.0 ST Array [transcript (gene) version], Affymetrix Human Genome U133A Array and Affymetrix Human Genome U133B Array platforms were provided by the illuminaHumanv4.db (Dunning et al.), hugene10sttranscriptcluster.db (MacDonald), hgu133a2.db (Carlson, a), and hgu133b.db (Carlson, b) packages respectively. The annotation packages and their respective data subsets are described in table 3.

Table 3: Annotation packages corresponding to data subsets.

Data subset	**Annotation package**
A1	hgu133a2.db
A2	hgu133a2.db
A3	hgu133b.db
A4	hgu133b.db
CC1	illuminaHumanv4.db
CC2	illuminaHumanv4.db
CC3	illuminaHumanv4.db
C1	hugene10sttranscriptcluster.db
GT1	hgu133a2.db

6.1.2 GO-term enrichment analysis

Differentially expressed genes were determined from the rows of the data subsets by `rowttests` function from genefilter package (Gentleman et al.). The cutoff P-values were different when performing enrichment analysis for biological process (BP) or molecular function (MF) GO-terms. They are listed in tables 4 and 5, for BP and MF analyses respectively. The P-values were determined in order to yield lists of terms which were able to be processed manually, therefore lists with possibly hundreds of terms were trimmed down by decreasing the P-value. Fold changes were analysed using `analyzeelbow` function from ELBOW package (Zhang et al., 2014), which contains a method for determining biologically significant fold changes.

Only expression values with significant fold change were considered in the GO-term enrichment analysis.

Table 4: P-value cutoffs for t-tests filtering the data for BP analysis.

Data subset	**Cutoff for BP analysis**
A1	1^{-4}
A2	1^{-5}
A3	1^{-3}
A4	1^{-4}
CC1	3^{-4}
CC2	5^{-5}
CC3	5^{-5}
C1	1^{-4}
GT1	5^{-3}

Table 5: P-value cutoffs for t-tests filtering the data for MF analysis.

Data subset	**Cutoff for MF analysis**
A1	5^{-4}
A2	1^{-4}
A3	5^{-4}
A4	1^{-4}
CC1	1^{-3}
CC2	1^{-4}
CC3	1^{-3}
C1	1^{-4}
GT1	5^{-3}

GO-term enrichment analysis was done with the GOstats package, utilizing the hypergeometric test (Falcon and Gentleman, 2007). As the sizes of the data sets varied, unique P-values were determined for each data subset. The ones chosen were set between 0.01 and 1^{-6} to reduce the chances of getting false negative or positive results. P-values for biological process (BP) and molecular function (MF) enrichment analysis are listed in tables 6 and 7. The results of gene enrichment analyses from different data subsets were compared with each other using `termLists` function (Appendix C), which counts the subsets where each GO-term was enriched.

Table 6: *P*-values for conducting enrichment analyses for BP terms and how many terms were found to be enriched.

Data subset	***P*-value for BP terms**
A1	1^{-5}
A2	1^{-6}
A3	5^{-4}
A4	1^{-3}
CC1	1^{-5}
CC2	2^{-5}
CC3	1^{-6}
C1	1^{-3}
GT1	1^{-4}

Table 7: *P*-values for conducting enrichment analyses for MF terms and how many terms were found to be enriched.

Data subset	***P*-value for MF terms**
A1	5^{-3}
A2	5^{-3}
A3	1^{-2}
A4	1^{-2}
CC1	1^{-3}
CC2	5^{-3}
CC3	1^{-4}
C1	5^{-3}
GT1	1^{-2}

6.1.3 The most differentially expressed genes and genes comprising selected GO-terms

The most differentially expressed genes in each data set were determined by using `rowttests` function from the genefilter package (Gentleman et al.), and setting the *P*-value cutoffs according to table 8. The probe lists were converted to gene names by using the respective annotation packages listed in table 3.

Table 8: P-values for finding the most differentially expressed genes, and the sizes of the resulting lists of genes.

Data subset	***P*-value**
A1	1^{-8}
A2	1^{-8}
A3	1^{-7}
A4	1^{-7}
CC1	1^{-5}
CC2	1^{-6}
CC3	1^{-6}
C1	1^{-6}
GT1	1^{-4}

Seven GO-terms listed in table 9 were selected for closer inspection. From those, the genes comprising each term were extracted by `extractGenes` function (Appendix D). The formed lists of genes were compared with the lists of the most differentially expressed genes to find out possible overlaps. This was done by `compareGeneLists` function (Appendix F). The fold change of each gene in different data subsets was found by `findFoldChange` function (Appendix G). Whether the gene had a homologue in *S. cerevisiae* was determined by manually searching Uniprot (Consortium, 2014).

Table 9: Selected GO-terms and their GO-IDs.

GO-term	**MF or BP**	**GO-ID**
NADH dehydrogenase activity	MF	GO:0003954
Oxidoreductase activity	MF	GO:0016491
Response to endoplasmic reticulum stress	BP	GO:0034976
Cellular response to unfolded protein	BP	GO:0034620
Protein targeting to ER	BP	GO:0045047
ER-associated ubiquitin-dependent protein catabolic process	BP	GO:0030433
Cell activation	BP	GO:0001775

6.2 Laboratory experiments

Four genes were chosen for knockout and overexpression experiments in antibody-secreting *S. cerevisiae*. The used strain was W303α strain (ATCC®208353TM, genotype: *MATa leu2-3,112 trp1-1 can1-100 ura3-1 ade2-1 his3-11,15*) having a DNA fragment expressing IgG integrated into HIS3-site by the use of pRS303N vector described in Taxis and Knop (2006). The IgG expression cassette has genes for the light and heavy chain encoding the monoclonal antibody C2B8 under GAL1-promoter. The chosen genes were Sec61 beta subunit (*SBH1*), glycerol-3-phosphate dehydrogenase 2 (mitochondrial) (*GPD2*), thioredoxin 1 (*TRX1*), and thioredoxin 2 (*TRX2*). Each gene was both knocked out and overexpressed in the yeast, and the effects were analyzed by enzyme-linked immunosorbent assay (ELISA).

6.2.1 Primers for the PCR steps

LoxP-kanMX-loxP marker cassettes (Güldener et al., 1996) were used for knockout experiments. 50 bp from both sides of each gene were combined with 20 bp from up- or downstream loxP site. The primers are described in table 10.

Table 10: Knockout primers. The bolded parts are loxP sites and the rest of the sequences are from either side of the genes.

Primer	Sequence	Tm (°C)
SBH1, forward	5'-ATG TCA AGC CCA ACT CCT CCA GGT GGT CAA CGT ACT TTG CAA AAG AGA AA **GTA CGC TGC AGG TCG ACA AC**-3'	91.0
SBH1, reverse	5'-TTA AAA TAA CTT ACC GGC AAC TTT AGA AAT AAC ATG TAA TGC AAC AAC AG **CCA CTA GTG GAT CTG ATA TC**-3'	81.4
GPD2, forward	5'-ATG CTT GCT GTC AGA AGA TTA ACA AGA TAC ACA TTC CTT AAG CGA ACG CA **GTA CGC TGC AGG TCG ACA AC**-3'	88.7
GPD2, reverse	5'-CTA TTC GTC ATC GAT GTC TAG CTC TTC AAT CAT CTC CGG TAG GTC TTC CA **CCA CTA GTG GAT CTG ATA TC**-3'	86.1
TRX1, forward	5'-ATG GTT ACT CAA TTC AAA ACT GCC AGC GAA TTC GAC TCT GCA ATT GCT CA **GTA CGC TGC AGG TCG ACA AC**-3'	90.9
TRX1, reverse	5'-TTA AGC ATT AGC AGC AAT GGC TTG CTT AAT AGC CGC TGG GTT GGC ACC AA **CCA CTA GTG GAT CTG ATA TC**-3'	89.1
TRX2, forward	5'-ATG GTC ACT CAA TTA AAA TCC GCT TCT GAA TAC GAC AGT GCT TTA GCA TC **GTA CGC TGC AGG TCG ACA AC**-3'	89.0
TRX2, reverse	5'-CTA TAC GTT GGA AGC AAT AGC TTG CTT GAT AGC AGC TGG GTT GGC ACC GA **CCA CTA GTG GAT CTG ATA TC**-3'	88.4

Overexpression primers were designed according to exonuclease and ligation-independent cloning (ELIC) protocol (Koskela and Frey, 2014). For genes *SBH1* and *GPD2*, the forward primer was designed to start at the gene promoter, and for both thioredoxins the forward primer started at the beginning of the gene. The vector used in overexpression experiments was pRS415-Tef, described in Mumberg et al. (1995). The primers are described in table 11.

Table 11: Overexpression primers. The 5' end of each primer is complementary to the plasmid and the 3' end is complementary to the gene or the promoter. Restriction enzymes are in the middle. The first Tm value is the Tm of the whole primer, and the second one is Tm of the part complementary to the gene or the promoter.

Primer	Sequence	Tm1 (°C)	Tm2 (°C)
5'-pAF5-SacI-SBH1-3'	5' -GGG AAC AAA AGC TG-GAG CTC-GTT CAT TAA ATT TCT A-3'	75.2	39.1
5'-pAF5-XhoI-SBH1-3'	5'-AAC TAA TTA CAT GA-CTC GAG-TTA AAA TAA CTTA CCG-3'	66.9	43.1
5'-pAF5-SacI-GPD2-3'	5' -GGG AAC AAA AGC TG-GAG CTC-CTA TTA TAG TGG GGA GAG-3'	77.6	48.9
5'-pAF5-XhoI-GPD2-3'	5'-AAC TAA TTA CAT GA-CTC GAG-CTA TTC GTC ATC GAT GTC-3'	74.2	54.2
5'-pAF5-SpeI-TRX1-3'	5'-CTA AGT TTT CTA GA-ACT AGT-ATG GTT ACT CAA TTC-3'	63.0	41.4
5'-pAF5-XhoI-TRX1-3'	5'-AAC TAA TTA CAT GA-CTC GAG-TTA AGC ATT AGC AGC-3'	70.1	47.6
5'-pAF5-SpeI-TRX2-3'	5'-CTA AGT TTT CTA G-ACT AGT-ATG GTC ACT CAA TTA-3'	64.0	42.7
5'-pAF5-XhoI-TRX2-3'	5'-AAC TAA TTA CAT GA-CTC GAG-CTA TAC GTT GGA AGC-3'	71.3	46.3

All primers were synthesized by Eurofins MWG Operon (Germany).

6.2.2 PCR and the preparation of DNA fragments

When preparing DNA fragments for knockout experiments, plasmid pUG6 carrying the loxP–kanMX–loxP disruption module (Güldener et al., 1996) was used as a template. In each of the four reaction mixes there were 3 mM plasmid, 2 mM

of both forward and reverse primers, 0.4 mM dNTPs, and 10 mM Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific) in Phusion HF Buffer (Thermo Fisher Scientific). The PCR programme is shown in table 12. After the final 10 minutes extension step, the PCR products were stored at 4 °C. After PCR, loading dye (Thermo Fisher Scientific) was added, and products were run on 0.7% agarose gel. Purification from gel was done with NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) according to the manufacturer's protocol, except the final elution which was done to 20 µl H_2O.

Table 12: PCR programme for preparing DNA fragments for knockouts. Lid temperature was 105 °C, and preheating was on.

Step	**Temperature (°C)**	**Time**	**Cycles**
1	98	30 s	
2	98	10 s	
3	54	30 s	
4	72	45 s	6 x to step 2
5	98	10 s	
6	68	30 s	
7	72	45 s	36 x to step 5
8	72	10 min	

For preparing DNA fragments for overexpression experiments, 0.04 mM yeast genomic DNA was used as template DNA. Otherwise the mixtures were same as in the knockouts described above. PCR programmes were different for *GPD2* and the other three genes, since the Tm values of the primers were different. For *GPD2*, the programme was the same as in table 12, except the temperature was 52 °C in step 3, and 77 °C in step 6. For the others, the programme was also the same as in table 12, except the temperature was 46 °C in step 3 and 67 °C in step 6, and steps 4 and 7 both lasted for 15 s.

After PCR, loading dye (6x DNA Loading Dye, Thermo Fisher Scientific) was added, and *GPD2* was run on 0.7% agarose gel, and the others were run on 1.2% agarose gel. Purification was done with NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) according to the manufacturer's protocol. The centrifuge used in all purifications was Eppendorf Centrifuge 5418. DNA concentrations were measured on Take3-platform with Eon Microplate Spectrophotometer (BioTek).

6.2.3 Plasmid generation and selection

Plasmid generation was done according to ELIC protocol (Koskela and Frey, 2014). Vector was digested in a mixture of 30 µl in 1 x FastDigest Green Buffer (Thermo Fisher Scientific), containing 5 µl pRS415-Tef vector, 1 FDU XhoI restriction enzyme (FastDigest, Thermo Fisher Scientific), and 1 FDU either SpeI (FastDigest, Thermo Fisher Scientific) or Ecl136II (FastDigest, Thermo Fisher Scientific) restriction enzyme. Mixtures were incubated overnight at +37 °C, and run on 0.7% agarose. Purification

was done with NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) according to the manufacturer's protocol.

The amount of insert was calculated according to equation

$$\frac{vector(ng) * insert(kb)}{vector(kb)} * 3 = insert(ng) \quad (2)$$

Vector amount was 80 ng for *GPD2* insert, and 100 ng for all others. *SBH1* and *GPD2* inserts were mixed with the plasmid digested with Ecl136II restriction enzyme, and the other two were mixed with the plasmid digested with SpeI restriction enzyme. H_2O was added for total volume of 10 µl. Transformation into chemical competent cells and screening of colonies were done as described in ELIC protocol (Koskela and Frey, 2014). Enzymes chosen for screening colonies for correct insert were Ecl136II and XhoI for *SBH1*, Eco31I for *GPD2*, EcoRI for *TRX1*, and SpeI and XhoI (all enzymes were FastDigest enzymes by Thermo Fisher Scientific) for *TRX2*, respectively. Test digestion mixtures were analyzed on 1% agarose.

6.2.4 Transformation

Both knockout cassettes and overexpression plasmids were transformed into YEK018 yeast by LiAc transformation from liquid culture. The yeast strain is described in the beginning of this section about laboratory experiments. Also pEK7 plasmids were transformed into knockout yeast strains after the confirmation of the knockout, so they could be cultivated in the same conditions as the overexpression yeast strains. This plasmid is like pRS415 plasmid (Mumberg et al., 1995), but lacks polylinker, promoter, and terminator removed with SacI/NaeI sites, thus it has only the empty vector backbone with LEU2 marker. Cells were grown overnight in 3 ml YPD, 30 °C, 220 RPM. OD600 values were measured with eppendorf BioPhotometer Plus machine, and samples were diluted to 0.5 OD600 (knockout cassettes) or 0.4 OD600 (overexpressions), and incubated at 30 °C, 220 RPM for three hours. Cells were harvested by centrifugation at 3900 RPM for 5 minutes. After resuspending the pellets in 50 ml sterile water the centrifugation was repeated. The pellets were resuspended in 500 µl solution containing 1 x TE and 0.1% LiAc, centrifuged for 3 minutes at 3600 RPM, and resuspended again in the same solution, the amount of which was 1/50 of the original culture volume.

100 µl of the harvested cells were mixed with either 500 ng of knockout cassettes or 600 ng of insertion plasmids. 30 ng denaturated salmon sperm was added as the carrier DNA along with 1 ml solution containing 1x TE, 0.1% LiAc, and 40% PEG, and the mixtures were incubated for 1 hour at 30 °C. After incubation, 120 µl DMSO was added, and cells were heat shocked at 42 °C for 10 minutes, cooled immediately in ice for 3 minutes, and centrifuged for 5 minutes at 3000 RPM. In overexpression experiments and when adding pEK7 plasmid to knockout strains, the pellets were resuspended in 100 µl sterile water, and plated on SD-Leu plates. When doing knockouts, the pellets were resuspended in 1 ml YPD and incubated for three hours at 30 °C, 220 RPM. The centrifugation was repeated and pellets were resuspended in 100 µl sterile water, and plated on YPD plates supplemented with

200 µg/ml G418. Further yeast cultivation is described in section 6.2.6. Centrifuge used for harvesting cells was Eppendorf Centrifuge 5810R, and the centrifuge used in other steps was Eppendorf Centrifuge 5418.

6.2.5 Verification of knockouts

After transformations, knockout yeasts were plated on YPD plates supplemented with 200 µg/ml G418, which is selective for kanMX marker in the knockout cassettes, and incubated at 30 °C for four days. In addition to the cultivation on selective media, the knockouts were colony-purified by selecting four colonies of each knockout yeast and plating them on new YPD plates supplemented with 200 µg/ml G418. These were incubated for two days at 30 °C. After storing the plates at 4 °C for 12 days, a colony from each sample on the new plates was transferred on a new YPD plates supplemented with 200 µg/ml G418 and incubated at 30 °C for three days. After incubation, one colony from each knockout yeas strain was chosen to be grown overnight at YPD (30 °C, 210 RPM) and stored at -80 °C.

6.2.6 Yeast cultivation and IgG expression

Overexpression yeast strains and knockout yeast strains containing the plasmid pEK7 were plated on SD-Leu plates, and incubated at 30 °C for two days. Two colonies from each sample were transferred to new SD-Leu plates that were incubated in 30 °C for four days. One colony from each yeast was selected and suspended into 1 x SD-Leu media containing 2% raffinose and 10 mM $NaPO_4$ as buffer. Cells were grown for 41 hours at 30 °C, 220 RPM. Normally they would have been incubated only for overnight, but there was not enough growth. These samples were used for ELISA.

OD600 values were measured with eppendorf BioPhotometer Plus machine, and samples were diluted to 0.5 OD600. Dilution mixture was the same as the SD-Leu media described above, supplemented with 0.05 mg/ml BSA (bovine serum albumin). Cells were grown on deep-well plate at 30 °C, 220 RPM for 5 hours 40 minutes. Protein expression was induced with different galactose concentrations (0.5%, 1%, 2%, and 4%), and samples were incubated for 24 h at 25 °C, 230 RPM. The cells were centrifuged at 3900 RPM for 10 minutes, OD600 values were measured with Eon Microplate Spectrophotometer (BioTek), and 285 µl supernatant from each sample was mixed with 15 µl 20 x PBT (PBS (135 mM NaCl, 2.5 mM KCl, 10 mM Na_2HPO_4, 1.75 mM KH_2PO_4) and 0.5% Tween-20), and stored in -22 °C.

6.2.7 ELISA

A 96-well plate was coated the day before by adding 38.5 µl per well of 4.2 µl of goat anti Human IgG (Fc specific) antibody (Sigma-Aldrich, Helsinki, Finland) in 20 ml 1 x PBS, and incubating the plate overnight at 7 °C with shaking. ELISA was conducted according to the following protocol:

Each well on the plate was washed 5 times with 200 µl PBT, and incubated with 200 µl PBT in room temperature, with shaking, for 45 minutes. The washing

was repeated, and samples along with standard series were added to each well. The volume of each sample and standard series specimen was 200 µl. The standard series consisted of human IgG standard antibody, obtained from Calbiochem (Merck Millipore, Billerica, USA), in PBT with the concentrations of 0, 0.5, 0.25, 0.125, 0.0625, 0.03125, 0.015625, and 0.0078125 mg/ml. The plate was incubated at +22 °C with shaking for 90 minutes, and washed 5 times with 200 µl PBT per well.

100 µl per wellof 1:4,000 dilution of goat anti-Human IgG (Fc specific)-peroxidase labelled antibody (Sigma-Aldrich, Helsinki, Finland) was added as detection antibody in PBT, with the total volume of 200 µl per well. The plate was incubated for 1 hour in room temperature, with shaking, and washed 5 times with 200 µl PBT per well. 80 µl of substrate consisting of 0.2 mg/ml *o*-phenylenediamine per well in 0.009 % H_2O_2 and 0.05 M phosphate-citrate buffer, was added to each well and incubated for 8 minutes. Reaction was stopped with 100 µl per well of 3 M H_2SO_4. The signal was measured by Synergy 2 plate reader (BioTek) and the data analysed with Gen5 2.01 software (BioTek). ELISA was conducted on Hamilton Microlab Star workstation (Hamilton Company), operated with the manufacturer's software. ELx405 Select Deep Well Washer (BioTek) was connected on the machine along with the above plate reader.

7 Results and discussion

7.1 Enriched GO-terms and genes associated with them

The enrichment analysis was evaluated for the overrepresentation of gene ontology (GO) terms, and conditional hypergeometric calculation was used. Analysis was conducted separately for each of the nine data subsets listed in table 2 and for both biological process (BP) and molecular function (MF) ontologies, thus 18 analyses were conducted in total. The background gene universe comprised of all the genes present in the respective microchip used in the original microarray experiment. The analyses resulted in total of 75 BP GO-terms and 67 MF GO-terms.

P-value cutoff parameters for filtering the data and conducting enrichment analyses can be found in tables 13, 14, 15, and 16 respectively. The chosen *P*-values were aimed to minimize both the false positive and false negative values. When filtering data, cutoff *P*-values ranged from 5^{-3} to 1^{-5}. The aim was to have a sufficient list of probes for the fold change analysis and for the further enrichment analysis while keeping the resulting lists short enough that the results of the enrichment analysis could be processed manually. The lower values were used for data subsets which had plenty of probes with significant change between the initial and final conditions. The aim was to filter the list and find the most significant ones from a larger gene list. The *P*-values for the enrichment analyses ranged from 1^{-2} to 1^{-6}. The values were chosen in order to gain lists of terms which were not too long to be processed manually. These values yielded lists of enriched terms ranging from 5 to 26 terms.

Table 13: *P*-value cutoffs for *t*-tests filtering the data for BP analysis.

Data subset	**Cutoff for BP analysis**	**Genes out of total**
A1	1^{-4}	1159/19915
A2	1^{-5}	721/19915
A3	1^{-3}	1057/15094
A4	1^{-4}	502/15094
CC1	3^{-4}	558/34298
CC2	5^{-5}	458/34298
CC3	5^{-5}	493/34298
C1	1^{-4}	505/19962
GT1	5^{-3}	340/19915

Table 14: *P*-value cutoffs for *t*-tests filtering the data for MF analysis.

Data subset	**Cutoff for MF analysis**	**Genes out of total**
A1	5^{-4}	1926/19915
A2	1^{-4}	1394/19915
A3	5^{-4}	803/15094
A4	1^{-4}	502/15094
CC1	1^{-3}	1208/34298
CC2	1^{-4}	682/34298
CC3	1^{-3}	2133/34298
C1	1^{-4}	505/19962
GT1	5^{-3}	340/19915

Table 15: *P*-values for conducting enrichment analyses for BP terms and how many terms were found to be enriched.

Data subset	***P*-value for BP terms**	**Amount of terms**
A1	1^{-5}	19
A2	1^{-6}	20
A3	5^{-4}	5
A4	1^{-3}	26
CC1	1^{-5}	10
CC2	2^{-5}	11
CC3	1^{-6}	13
C1	1^{-3}	8
GT1	1^{-4}	8

Table 16: *P*-values for conducting enrichment analyses for MF terms and how many terms were found to be enriched.

Data subset	***P*-value for MF terms**	**Amount of terms**
A1	5^{-3}	17
A2	5^{-3}	17
A3	1^{-2}	5
A4	1^{-2}	11
CC1	1^{-3}	5
CC2	5^{-3}	12
CC3	1^{-4}	8
C1	5^{-3}	7
GT1	1^{-2}	5

Complete lists of the enriched terms in all data subsets are listed in appendix H, and the terms enriched in more than one data subset are listed in tables 17 and 18.

The enriched GO-terms showed a lot of variation between data subsets. 21 out of 75 BP terms and 11 out of 67 MF terms were enriched in more than one data subset. These BP terms were all associated with cell activation or B cell development into plasma cells. The MF terms enriched in multiple data subsets comprised of broad terms, such as enzyme binding, and terms associated with B cell development, like MHC class II receptor activity. Of the data subsets, A2 had most BP (15 terms) and MF (9 terms) terms common with at least one other subset. CC3 did not have any BP terms common with the other subsets, and A3 had no MF terms common with the others. The origin of the cells does not seem to affect how many of the terms were common with the other subsets, as both subsets A2 and CC3 comprised cells harvested from bone marrow.

Terms enriched in the data of (Cocco et al., 2012) can be found in their supplemental data. 18/21 of the BP terms enriched in multiple data subsets were also enriched in the study by Cocco et al. (2012). The missing terms were lymphocyte proliferation, regulation of B cell proliferation, and regulation of response to stimulus. The first two ones were not enriched in any of the data subsets formed from the data of Cocco et al. (2012), and the third one had its child terms, negative regulation of response to stimulus and positive regulation of response to stimulus, found to be enriched. 5/11 of the MF terms enriched in multiple data sets were found to be enriched in the study by Cocco et al. (2012). Of the terms not found to be enriched by Cocco et al. (2012), poly(A) RNA binding and phosphatidylinositol 3-kinase binding were found in data subsets from Cocco et al. (2012). This is most likely due to different algorithms used in the enrichment analysis. The other four MF terms not found to be enriched in the data of Cocco et al. (2012) were histone binding, nucleic acid binding, chromatin binding, and CD8 receptor binding.

Romijn et al. (2005) studied the expression levels of developing plasma cells. They were able to quantify the expression level 234 proteins at every one of the total of five inspected time points during plasma cell development.They did not utilize GO annotation, but did annotate the proteins according to their function into 11 distinct categories, plus 12th one for all other functions. The categories were immune system, ER targeting and folding, thiol redox balance, cytosolic and mitochondrial chaperones, translation, degradation, signaling, membrane transport, metabolism, cytoskeleton, and cell cycle. Almost all of the BP terms enriched in multiple data subsets can be grouped into the immune system category, except the broad terms regulation of response to stimulus, cell activation, positive regulation of cell activation, and regulation of cell activation, which did not fit into any of the categories formed. B cell receptor signaling pathway is also involved in signaling. Five of the MF terms enriched in multiple data subsets were involved with cell cycle (histone binding, nucleic acid binding, chromatin binding, poly(A) RNA binding, phosphatidylinositol 3-kinase binding), three with immune system (MHC class II receptor activity, CD8 receptor binding, antigen binding), one with translation (structural constituent of ribosome), one with signaling (MHC class II receptor activity), and two were uncategorized (enzyme binding and protein kinase binding). Except the very broad terms, all of the terms enriched in multiple data subsets can be assigned into the categories Romijn et al. (2005) found enriched, which supports that these terms are indeed enriched in

plasma cells.

van Anken et al. (2003) studied the protein expression during plasma cell development. They divided the enriched proteins into seven functional categories, cytoskeleton, cytosolic chaperones, mitochondrial chaperones, metabolism, redox balance, ER resident proteins, and immune response. As stated above, most of the BP terms and three of the MF terms enriched in multiple data subsets were involved with immune response. The other terms enriched in multiple data subsets do not clearly fall into any of the categories formed by van Anken et al. (2003), which in part shows how data from proteomic study like the one by van Anken et al. (2003) does not necessarily correlate with data from transcriptomics studies, as stated in section 4.2.

Table 17: Enriched BP terms that appear in more than one data subset, and their respective data subsets.

Biological process term	**Amount**	**Subsets**
immune response	6	CC1, CC2, C1, GT1, A1, A2
immune response-activating cell surface receptor signaling pathway	6	CC1, CC2, GT1, A1, A2, A4
B cell activation	5	CC1, CC2, A1, A2, A4
activation of immune response	4	CC1, A1, A2, A4
immune response-regulating signaling pathway	4	CC1, A1, A2, A4
positive regulation of immune system process	4	CC1, C1, A2, A4
B cell receptor signaling pathway	3	CC1, A1, A2
positive regulation of lymphocyte activation	3	CC1, GT1, A2
regulation of response to stimulus	3	CC1, CC2, A1
regulation of leukocyte activation	3	CC2, GT1, A2
regulation of B cell proliferation	3	C1, A1, A2
T cell activation	3	GT1, A2, A4
leukocyte activation	2	CC1, A4
cell activation	2	CC2, A3
lymphocyte activation	2	CC2, A3
positive regulation of cell activation	2	GT1, A2
regulation of lymphocyte activation	2	A1, A4
regulation of cell activation	2	A1, A4
lymphocyte proliferation	2	A1, A2
immune effector process	2	A1, A2
lymphocyte differentiation	2	A2, A4

Table 18: Enriched MF terms that appear in more than one data subset, and their respective data subsets.

Molecular function term	**Amount**	**Subsets**
enzyme binding	4	CC1, CC2, C1, A2
antigen binding	4	CC2, GT1, A1, A4
poly(A) RNA binding	3	CC1, A1, A2
protein kinase binding	3	CC1, CC2, A2
phosphatidylinositol 3-kinase binding	3	CC2, A1, A2
MHC class II receptor activity	3	CC3, GT1, A1
histone binding	3	A1, A2, A4
structural constituent of ribosome	2	CC1, A2
nucleic acid binding	2	A1, A2
chromatin binding	2	A1, A2
CD8 receptor binding	2	A1, A2

Seven terms out of the 142 enriched BP and MF terms were chosen for closer inspection on the basis of prior biological knowledge and the cell biology of the yeast. The chosen terms were NADH dehydrogenase activity, oxidoreductase activity, response to endoplasmic reticulum stress, cellular response to unfolded protein, protein targeting to ER, ER-associated ubiquitin-dependent protein catabolic process, and cell activation. NADH dehydrogenase activity and oxidoreductase activity are MF terms, the others are BP terms. Out of these seven terms only ER-associated ubiquitin-dependent protein catabolic process was not enriched in the study by Cocco et al. (2012).

As stated in sections 2 and 3, the ability to secrete proteins efficiently is vital for both yeasts used as cell factories and plasma cells. The GO-terms 'response to endoplasmic reticulum stress', 'ER-associated ubiquitin-dependent protein catabolic process', 'cellular response to unfolded protein', and 'protein targeting to ER' are all associated with the secretory pathway. 'Oxidoreductase activity' and its child term (Ashburner et al., 2000) 'NADH dehydrogenase activity' were found interesting, since during the production of antibodies the increased disulphide bond formation induces pressure on the redox conditions of the cell (Masciarelli and Sitia, 2008). 'Cell activation' was chosen because it was thought to comprise genes related to active cell metabolism in general.

126 genes in total were found to be associated with the chosen GO-terms. In order to see which of them contributed the most to plasma cell development, they were compared with the lists of the most differentially expressed genes in all separate data subsets. The P-value cutoffs for selecting the most differentially expressed genes are listed in table 8. The amount of genes yielded by the P-values ranged from 10 to 72 for different data subsets, which was deemed suitable for minimizing both false negative and false positive results. The P-values and the sizes of the gene lists yielded are listed in table 19. The most differentially expressed genes of each data subset are listed in appendix I. The most overlap was found among the genes associated with cell activation: 22 out of 53 genes were also identified as

the most differentially expressed ones. The smallest overlap was among the genes associated with NADH dehydrogenase activity, of which none were identified as the most differentially expressed genes.

Table 19: P-values for finding the most differentially expressed genes, and the sizes of the resulting lists of genes.

Data subset	***P*-value**	**Size of the list**
A1	1^{-8}	32
A2	1^{-8}	72
A3	1^{-7}	29
A4	1^{-7}	46
CC1	1^{-5}	17
CC2	1^{-6}	39
CC3	1^{-6}	19
C1	1^{-6}	55
GT1	1^{-4}	10

Out of the 126 genes associated with the chosen GO-terms, 24 had a *S. cerevisiae* homologue, and were therefore considered for the laboratory experiments. These 24 genes are listed in table 20 along with their mean fold changes and standard deviations (if applicable) across the data subsets. Nearly all of the genes were upregulated. Half of the genes were associated with oxidoreductase activity, nine were ribosomal proteins associated with protein targeting to ER, two were associated with both cellular response to unfolded protein and response to endoplasmic reticulum stress, and one was associated with ER-associated ubiquitin-dependent protein catabolic process.

The genes selected for *in vivo* testing were Sec61 beta subunit (*SBH1*), glycerol-3-phosphate dehydrogenase 2 (mitochondrial) (*GPD2*) and thioredoxin (*TRX*). *SBH1* was chosen since it is involved in the uptake of the nascent polypeptide chain into the lumen of the ER (Toikkanen et al., 1996), and therefore its overexpression could lead into improved rate of polypeptide chain uptake into the ER, resulting in increased rate of antibody secretion.

Since there are two cytoplasmic thioredoxins, thioredoxin 1 (*TRX1*) and thioredoxin2 (*TRX2*), in *S. cerevisiae*, they were both used in the *in vivo* testing. Thioredoxins act as hydrogen donors in redox reactions, and besides regulating the redox balance in the cytoplasm, they are involved in multiple cellular processes, including protein folding. Antibodies are rich in disulphide bonds, which increase their stability, therefore the ER needs to have a large oxidative capacity to meet the need of rapid introduction of disulphide bonds into antibodies. Disulphide bond formation is based on redox reactions, where two cysteines are oxidized by removing two electrons forming a covalent bond. During antibody production and expansion of the ER, an excess of electrons is produced due to the disulphide bond formation. This induces pressure on the redox conditions of the cell, and so additional buffer capacity can be needed the cytosol and mitochondria to maintain the redox homeostasis in these

compartments (Masciarelli and Sitia, 2008; Romijn et al., 2005; Trotter and Grant, 2002). *GPD2* oxidizes NADH, and is thus also involved in the redox homeostasis maintenance (Vemuri et al., 2007). It was also deemed interesting since it was the only downregulated gene which did not encode ribosomal protein subunit and which homologue in *S. cerevisiae* was not putative. It catalyses the production of glycerol under anaerobic growth conditions, and controls the rate of glycerol formation in *S. cerevisiae* (Cronwright et al., 2002; Hubmann et al., 2011).

According to the findings of van Anken et al. (2003), proteins functioning in redox balance, like the thioredoxins and glycerol-3-phosphate dehydrogenase, and ER resident proteins, such as Sec61 beta subunit, are enriched in the development of plasma cells, which supports their choice for the experiments *in vivo*.

Table 20: Original genes of the selected GO-terms which have a *S. cerevisiae* homologue. Two of the genes are associated with two different GO-terms. Mean is the mean of the fold changes in all data subsets, and SD is the standard deviation. The genes which have no value for standard deviation were present in only one data subset.

GO term	**Gene**	**Mean**	**SD**
Cellular response to unfolded protein / Response to endoplasmic reticulum stress	X-box binding protein 1	4.73	0.76
Cellular response to unfolded protein / Response to endoplasmic reticulum stress	alanyl-tRNA synthetase	2.65	0.51
ER-associated ubiquitin-dependent protein catabolic process	Sec61 beta subunit	2.68	0.50
Protein targeting to ER	ribosomal protein L12	-0.28	
Protein targeting to ER	ribosomal protein L15	-0.78	0.71
Protein targeting to ER	ribosomal protein L22	-0.56	0.16
Protein targeting to ER	ribosomal protein L27	-0.31	
Protein targeting to ER	ribosomal protein L38	-0.28	
Protein targeting to ER	ribosomal protein S2	-0.21	
Protein targeting to ER	ribosomal protein S23	-0.20	
Protein targeting to ER	ribosomal protein S27	-0.28	0.04
Protein targeting to ER	ribosomal protein S29	-0.22	
Oxidoreductase activity	saccharopine dehydrogenase (putative)	-0.15	
Oxidoreductase activity	pterin-4 alpha-carbinolamine dehydratase/dimerization cofactor of hepatocyte nuclear factor 1 alpha	1.38	
Oxidoreductase activity	glutaredoxin 2	0.99	
Oxidoreductase activity	thioredoxin	2.13	0.30
Oxidoreductase activity	isocitrate dehydrogenase 2 (NADP+), mitochondrial	2.56	0.43
Oxidoreductase activity	ubiquinol-cytochrome c reductase, complex III subunit VII, 9.5kDa	2.03	
Oxidoreductase activity	glycerol-3-phosphate dehydrogenase 2 (mitochondrial)	-0.15	
Oxidoreductase activity	protoporphyrinogen oxidase	0.74	
Oxidoreductase activity	methylsterol monooxygenase 1	1.64	
Oxidoreductase activity	cytochrome c oxidase subunit Va	1.41	
Oxidoreductase activity	phosphoglycerate dehydrogenase	2.89	
Oxidoreductase activity	ferric-chelate reductase 1	1.34	

7.2 Plasmids and yeast strains generated

The plasmids used in the overexpression experiments were generated by isolating the used genes and promoters using PCR, as described in section 6.2.2, and inserting them into expression plasmids according to the ELIC protocol (Koskela and Frey, 2014), as described in section 6.2.3. Genes *SBH1* and *GPD2* were expressed under their respective endogenous promoters. The two thioredoxins were expressed under the *TEF* promoter, since thioredoxins are highly regulated, so their respective promoters would probably include elements which would inactivate the genes in various conditions. Thus to ensure their overexpression, the genes *TRX1* and *TRX2* were expressed under the strong, constitutive *TEF* promoter. The plasmids generated in this study are mapped in pictures 10, 11, 12, and 13.

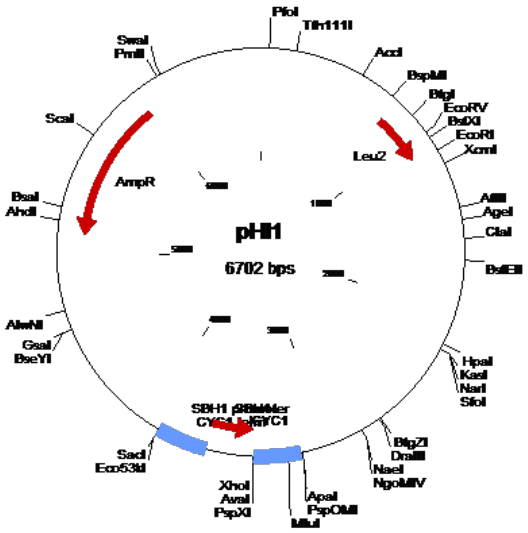

Figure 10: Plasmid pHI1, which includes the gene and promoter of *SBH1*, and *LEU2* marker.

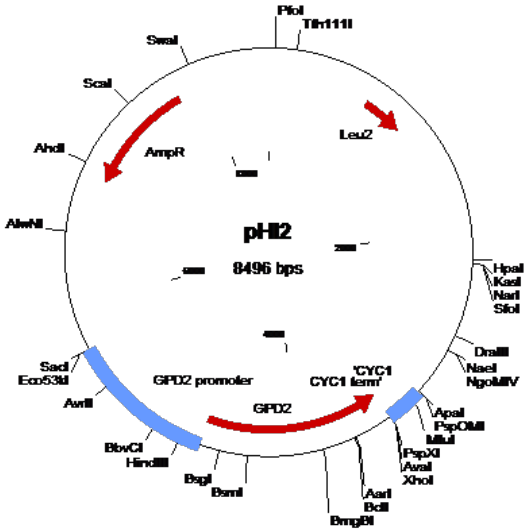

Figure 11: Plasmid pHI2, which includes the gene and promoter of *GPD2*, and *LEU2* marker.

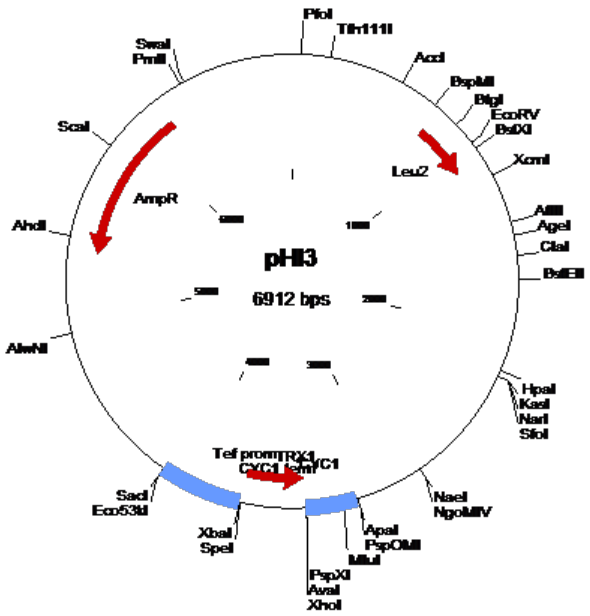

Figure 12: Plasmid pHI3, which includes the gene *TRX1*, under *TEF* promoter and *LEU2* marker.

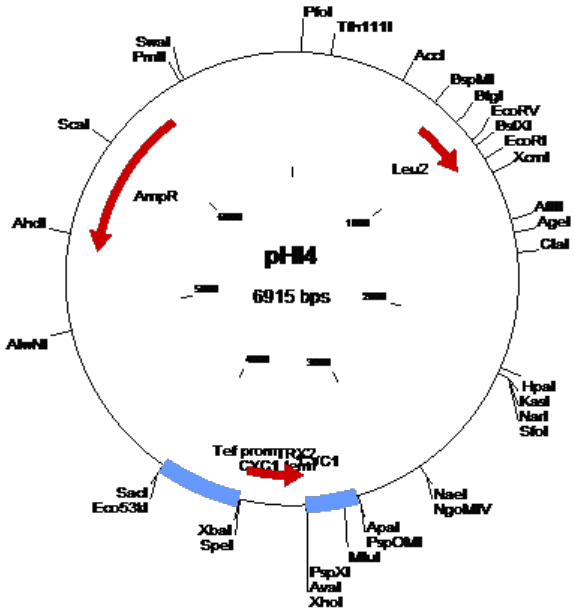

Figure 13: Plasmid pHI4, which includes the gene *TRX2* under *TEF* promoter, and *LEU2* marker.

Plasmid test digestion for pHI1, pHI3, and pHI4 is shown in figure 14. Since the test digestion was unclear for plasmid pHI2, new colonies were screened and test digested. The digestion is shown in picture 15. Protocol for test digestion is described in section 6.2.3.

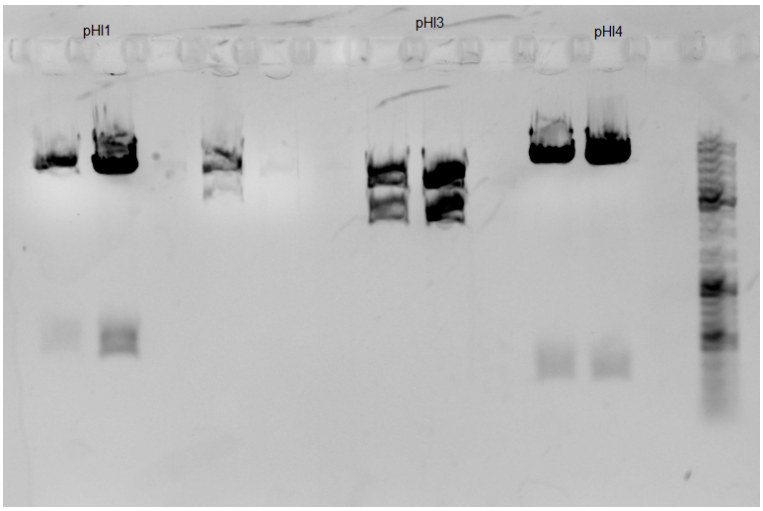

Figure 14: Plasmid test digestion run on 1% agarose. The restriction enzymes cut plasmid pHI1 into fragments of about 500 bp and 6000 bp, plasmid pHI3 into fragments of about 2500 bp and 3500 bp, and plasmid pHI4 into fragments of about 300 and 6500 bp. Two colonies of each plasmid were tested for the correct insert, and for pHI1, pHI3, and pHI4, the one on the right was chosen, respectively.

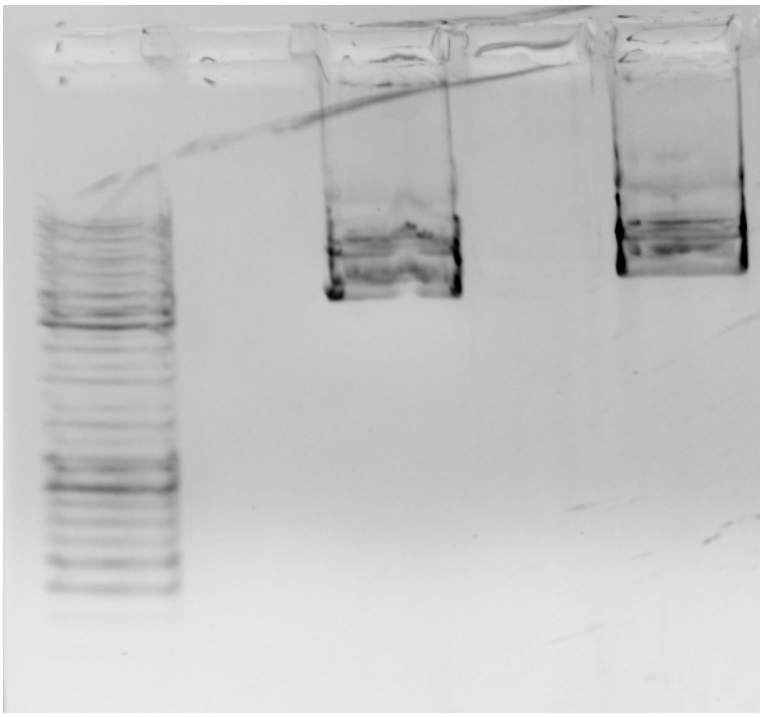

Figure 15: Plasmid test digestion run on 1% agarose. The chosen one is on the left, and it is cut into one fragment of about 8000 bp.

The deletion strains were generated using homologous recombination with PCR generated knockout cassettes, described in section 6.2.2. The cassettes, as well as

the overexpression plasmids and pEK7 plasmids, were transformed into yeast by LiAc transformation from liquid culture, according to the protocol in section 6.2.4. The knockout yeasts YHI001-YHI004 were grown on YPD plates supplemented with G418, which is selective for the kanMX marker in the knockout cassettes, as described in section 6.2.5. Yeast strains YHI005-YHI013 were cultivated on SD-Leu plates, as described in section 6.2.6. The yeast strains created in this study are listed in table 21. All yeast strains are based on YEK018 yeast strain, which is described in section 6.2. The plasmid pEK7 is described in section 6.2.4.

Table 21: Yeast strains generated in this thesis.

Yeast strain	Description
YHI001	YEK018Δ*sbh1*
YHI002	YEK018Δ*gpd2*
YHI003	YEK018Δ*trx1*
YHI004	YEK018Δ*trx2*
YHI005	YEK018 with plasmid pHI1
YHI006	YEK018 with plasmid pHI2
YHI007	YEK018 with plasmid pHI3
YHI008	YEK018 with plasmid pHI4
YHI009	YHI001 with plasmid pEK7
YHI010	YHI002 with plasmid pEK7
YHI011	YHI003 with plasmid pEK7
YHI012	YHI004 with plasmid pEK7
YHI013	YEK018 with plasmid pEK7

7.3 Testing *in vivo*

The genes *SBH1*, *GPD2*, *TRX1*, and *TRX2* were each overexpressed and knocked out in antibody-secreting *S. cerevisiae*. The antibody secretion was induced by galactose at four different concentrations; 0.5%, 1%, 2%, and 4%. The effects of the overexpressions and knockouts on the secretion of the antibody were analysed by ELISA. There were three rounds of ELISA, the first one had one plate, and the two other were both conducted with two plates, thus in total there were five plates. On one plate, there were two samples of each mutant yeast and two antibody-secreting control yeasts per each galactose concentration.

Overall there was a lot of variation between different rounds of ELISA in the measured levels of antibody. The first round was done directly after the transformation, so the yeasts had not been frozen. The last two rounds were done after re-striking the frozen yeasts on new SD-Leu plates. In the first round of ELISA, the magnitude of antibody concentration was $10^{-2} mg/ml$ across all specimens, whereas in the last round the magnitude was $10^{-3} mg/ml$. In the second round, i.e. plates 2 and 3, the magnitude was $10^{-2} mg/ml$ except the yeasts grown on 4% galactose concentration, which the magnitude was $10^{-3} mg/ml$. Thus the freezing hindered the secretion of the antibody. Freezing did not affect the OD600 values. The OD600 values did not

correlate with whether the yeast strain was knock-out or overexpression yeast. In order to minimize the bias arising from the differing antibody concentrations, four diagrams were drawn:

- Concentration of antibody
- Antibody concentration relative to OD600 value
- Concentration of antibody relative to control yeast
- Concentration of antibody / OD600 value relative to control yeast

In the last two diagrams, the mean values of the control yeasts of each plate and galactose concentration were set to 1, and the mutant yeast were compared to the control yeasts on the same plate and the same galactose concentration, i.e. the yeasts on plate one grown on 0.5% galactose were compared to the mean of the two control yeasts grown on 0.5% on that plate. In all diagrams, cyan is yeast strain grown on 0.5% galactose, magenta on 1% galactose, green on 2% galactose, and yellow is yeast strain grown on 4% galactose.

7.3.1 Sec61 beta subunit

SBH1 codes the beta subunit of the Sec61 complex involved in the uptake of the nascent polypeptide chain into the lumen of the ER. The hypothesis was that its overexpression would improve the uptake of the antibody into the ER, and hence enhance the antibody secretion, but this did not happen. The bottleneck of the antibody production therefore does not reside in the uptake of the nascent peptide chains into the ER, or the uptake occurs via the Ssh1p complex described by Finke et al. (1996). At the lowest galactose concentration (0.5%), the antibody secretion was hindered by the overexpression of *SBH1*, but however at the 1% galactose concentration the overexpression yeasts actually secreted more antibodies than their control counterparts. This could be due because the bottleneck does not lie in the step affected by the *SBH1* gene.

The knockout didn't effect the antibody secretion. This could mean that the nascent polypeptide chain can be translocated into the ER via the Ssh1p complex. The translocation into the ER could be further researched by the deletion and overexpression of *SBH2*, which is the counterpart of *SBH1* in the Ssh1p complex (Finke et al., 1996). If its overexpression enhances the antibody secretion and knockout hinders the secretion, it would indicate that the antibody is translocated into the ER by the Ssh1p complex.

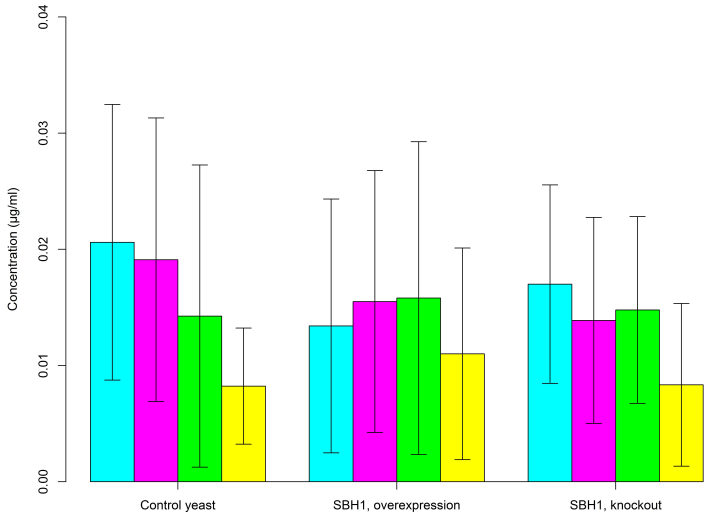

Figure 16: The antibody concentration ($\mu g/ml$) of control and both the overexpression and knockout *SBH1* yeasts. The variety between different plates was large, hence the big error margins.

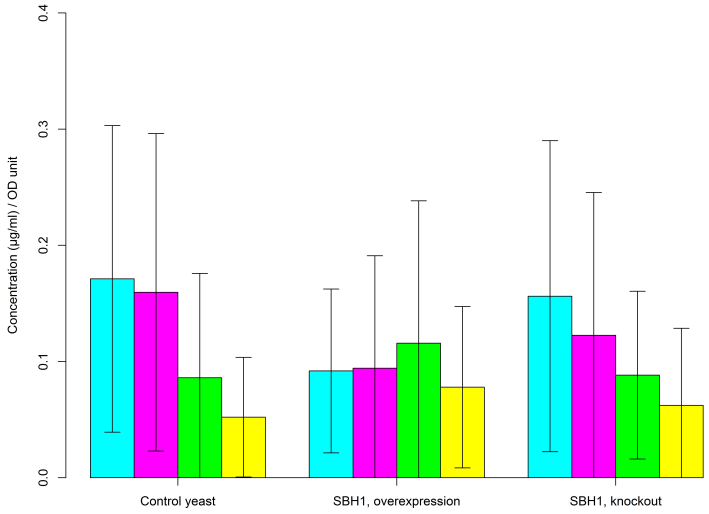

Figure 17: The antibody concentration ($\mu/ml)/ODunit$) of control and both the overexpression and knockout *SBH1* yeasts. The variety between different plates was large, hence the big error margins. The overexpression of *SBH1* lowered the antibody secretion at the two lowest galactose concentrations. Knockout did not affect the antibody secretion.

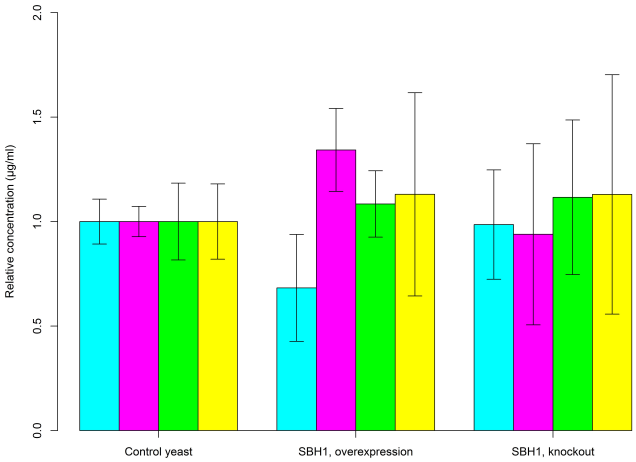

Figure 18: The relative antibody concentration of both the overexpression and knockout *SBH1* yeasts compared to control yeasts on each respective plate. When comparing the antibody concentration to the control yeasts on the same plate, the knockout yeasts didn't show any significant change in the level of secreted antibody. The overexpression yeasts had lower antibody secretion at 0.5% galactose, but higher at 1% galactose, and no significant changes at the two highest galactose consentrations.

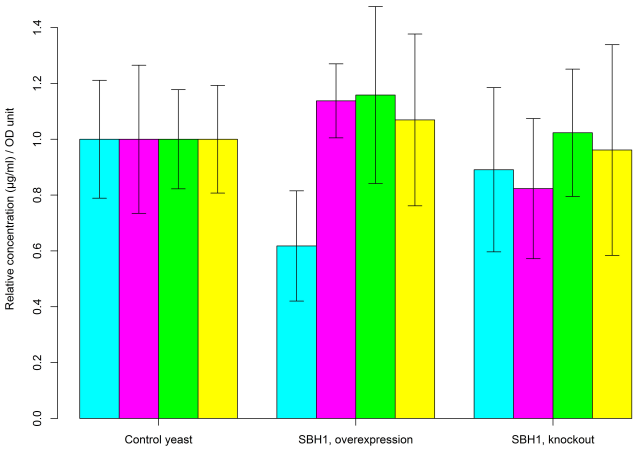

Figure 19: The relative antibody concentration per OD unit of both the overexpression and knockout *SBH1* yeasts compared to control yeasts on each respective plate. When compared to the control yeast on the same plate, the relative antibody concentration was significantly lower in the overexpression yeast at 0.5% galactose.

7.3.2 Glycerol-3-phosphate dehydrogenase 2 (mitochondrial)

The expression of glycerol-3-phosphate dehydrogenase 2, GPD2, was downregulated in plasma cells. In *S. cerevisiae*, it is the rate-controlling enzyme of the glycerol formation pathway (Cronwright et al., 2002).

The overexpression of *GPD2* slightly hindered the antibody secretion. It has been noted by Cronwright et al. (2002) that the overexpression of *GPD2* in *S. cerevisiae* may reduce the concentration of ATP by redirecting carbon towards glycerol formation pathway and away from the ATP-generating pathway. This could also cause the lower secretion of antibody.

The knockout of *GPD2* increased the relative antibody secretion, especially at the two highest galactose concentrations, 2% and 4%. This could be due to higher ATP concentration.

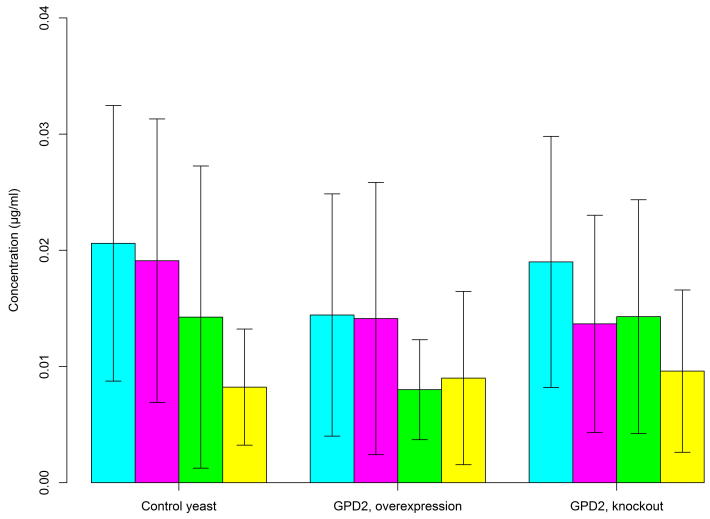

Figure 20: The antibody concentration ($\mu g/ml$) of control and both the overexpression and knockout *GPD2* yeasts. The variety between different plates was large, hence the big error margins. The overexpression of *GPD2* lowered the antibody concentration.

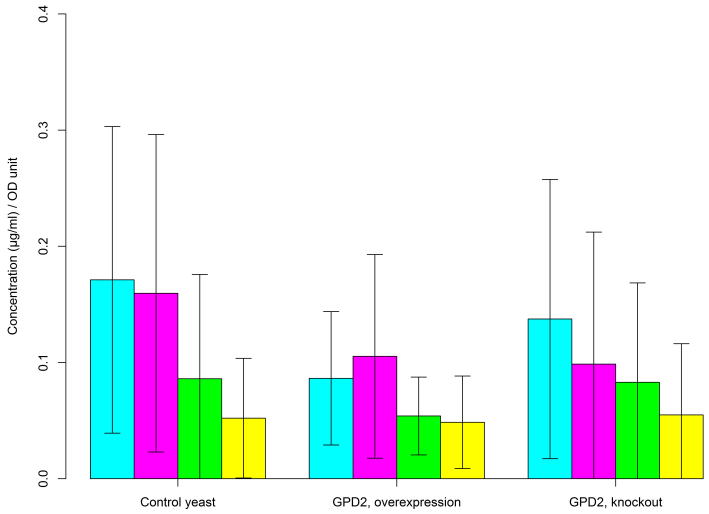

Figure 21: The antibody concentration $((\mu/ml)/ODunit)$ of control and both the overexpression and knockout *GPD2* yeasts. The variety between different plates was large, hence the big error margins, especially in the knockout yeasts. The antibody concentration per OD unit was lower in the overexpression yeasts than in the control yeasts.

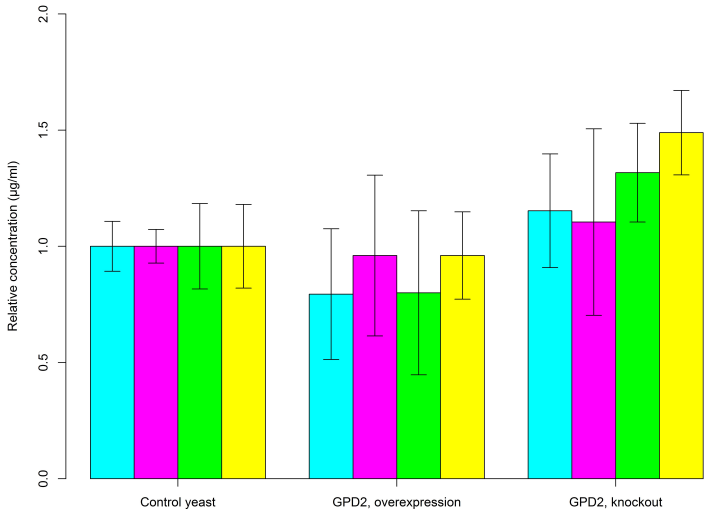

Figure 22: The relative antibody concentration of both the overexpression and knockout *GPD2* yeasts compared to control yeasts on each respective plate. The antibody concentration compared to control yeasts was highest for knockout yeasts at the two highest galactose concentrations. The overexpression of *GPD2* hindered the antibody secretion.

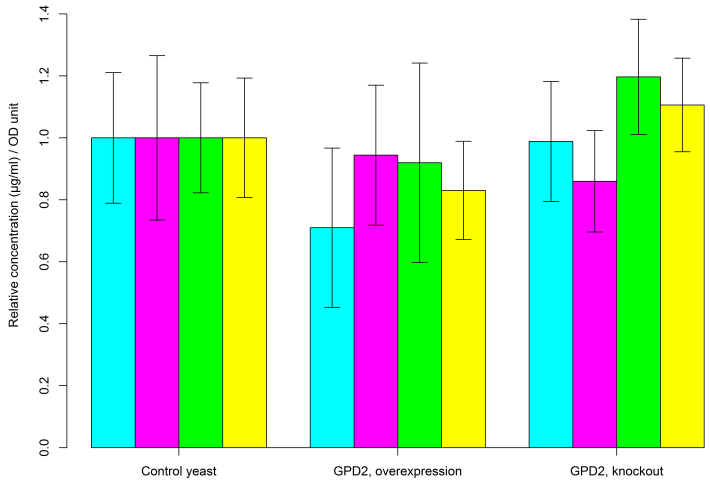

Figure 23: The relative antibody concentration per OD unit of both the overexpression and knockout *GPD2* yeasts compared to control yeasts on each respective plate. The antibody concentration per OD unit relative to control yeasts was lowest for the overexpression yeasts at the lowest galactose concentration, and the highest for the knockout yeasts at the two highest galactose concentrations.

7.3.3 Thioredoxin 1 and thioredoxin 2

There are two cytosolic thioredoxins in *S. cerevisiae*, which were studied separately. Thioredoxins are required in the maintenance of the redox homeostasis in the cell (Trotter and Grant, 2002). They also facilitate protein folding and enhance other molecular chaperones in refolding activity. The thioredoxin fold of thioredoxins can induce conformational changes in target disulphide by binding to target proteins through backbone-backbone hydrogen bonds. This is followed by a nucleophilic attack by a Cys-residue located in the N terminus of the fold.As a result, a mixed disulfide covalent bond is generated. Thioredoxins also aid the folding activities of other molecular chaperones (Berndt et al., 2008).

The overexpression of *TRX1* hindered the secretion of the antibody, but the overexpression of *TRX2* increased it. It has been noted by Garrido and Grant (2002) that *TRX1* could have an auxiliary role to *TRX2*. The expression levels of the two thioredoxins could be dependent of each other, and thus the overexpression of *TRX1*, inhibits the expression of the more effective *TRX2*. As the overexpression of *TRX2* increased the antibody secretion, it seems that the bottleneck in the production of antibody in *S. cerevisiae* lies in the protein folding step, and the thioredoxin could facilitate protein folding or help the maintenance of the redox homeostasis.

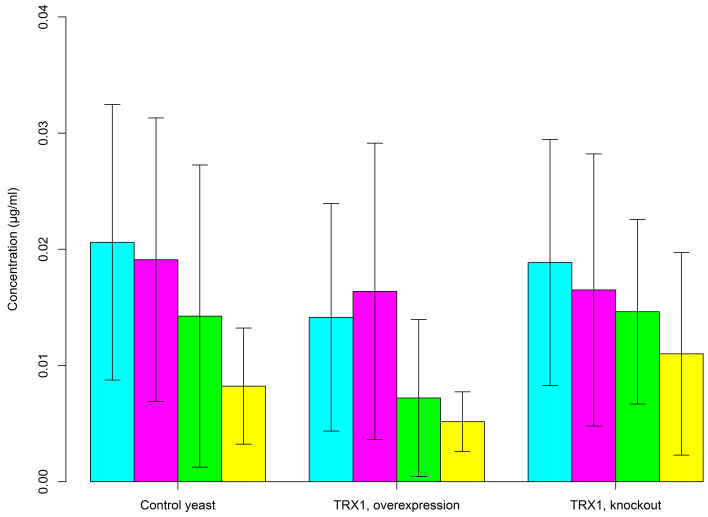

Figure 24: The antibody concentration ($\mu g/ml$) of control and both the overexpression and overexpression and knockout *TRX1* yeasts. The variety between different plates was large, hence the big error margins. The antibody secretion was lowered in the overexpression yeasts, especially at the two highest galactose concentrations. Knockout did not seem to affect the antibody secretion.

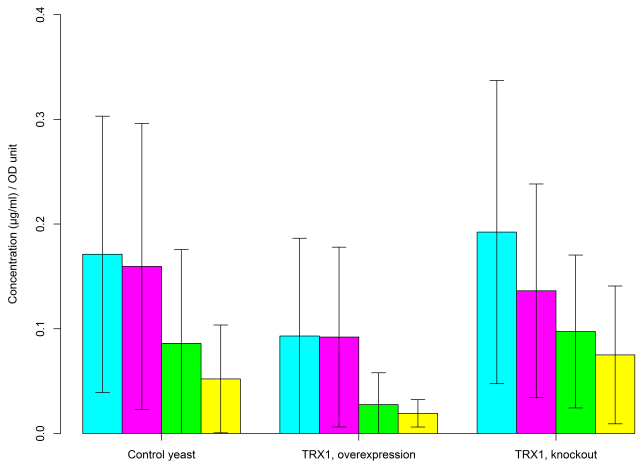

Figure 25: The antibody concentration ($(\mu/ml)/ODunit$) of control and both the overexpression and knockout *TRX1* yeasts. The variety between different plates was large, hence the big error margins. The overexpression of *TRX1* lowered the antibody concentration relative to OD units. Knockout didn't affect the antibody secretion per OD unit.

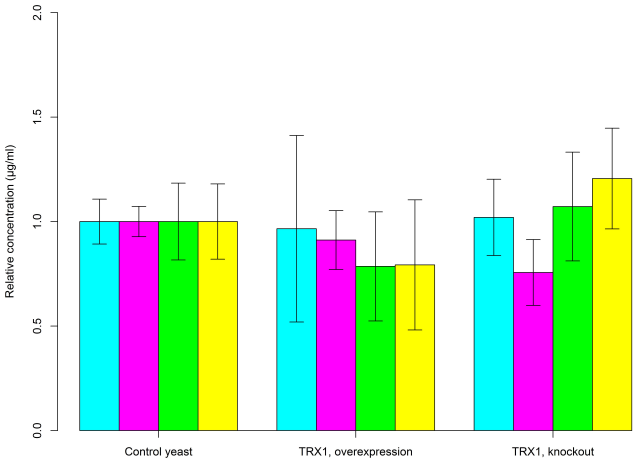

Figure 26: The relative antibody concentration of both the overexpression and knockout *TRX1* yeasts compared to control yeasts on each respective plate. The overexpression of *TRX1* lowered the antibody secretion relative to control yeast, however the error margins are large, especially for the yeasts grown on 0.5% galactose. The knockout seemed to lower the antibody secretion on 1% galactose, but enhance it on other concentrations.

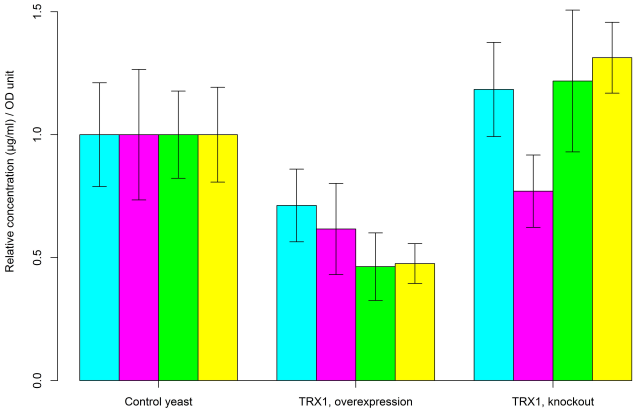

Figure 27: The relative antibody concentration per OD unit of both the overexpression and knockout *TRX1* yeasts compared to control yeasts on each respective plate. The overexpression of *TRX1* lowered the antibody concentration per OD unit relative to control yeast. The antibody secretion was hindered on all galactose concentrations. The knockout lowered the relative antibody concentration per OD unit on 1% galactose, but increased it on the other three galactose concentrations.

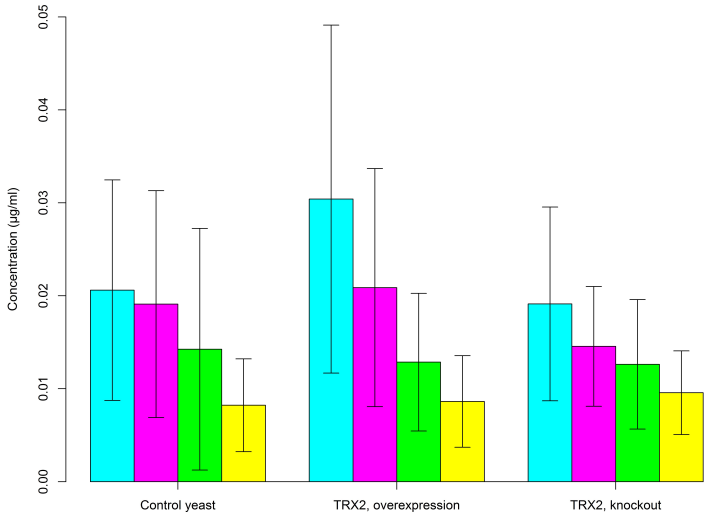

Figure 28: The antibody concentration ($\mu g/ml$) of control and both the overexpression and overexpression and knockout *TRX2* yeasts. The variety between different plates was large, and thus the error margins are large especially on the overexpression yeasts grown on 0.5% galactose. The overexpression of *TRX2* increased the antibody concentration, and the knockout slightly hindered the secretion.

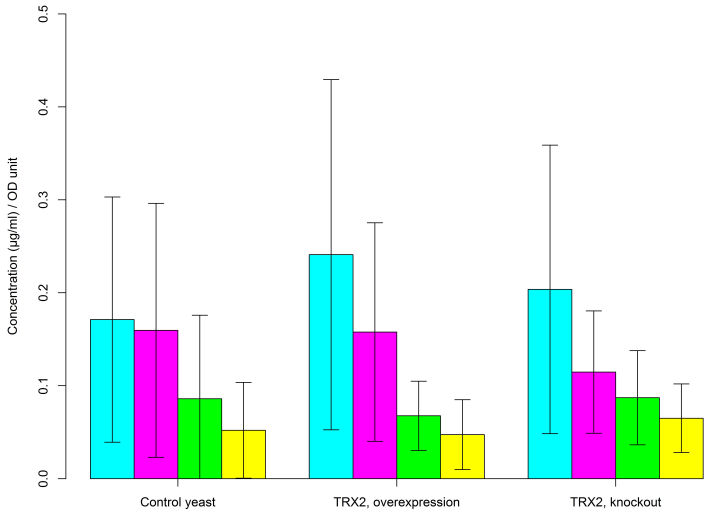

Figure 29: The antibody concentration ($(\mu/ml)/ODunit$) of control and both the overexpression and knockout *TRX2* yeasts. The variety between different plates was large, hence the big error margins. The overexpression of *TRX2* increased the antibody secretion relative to OD unit on the 0.5% galactose concentration.

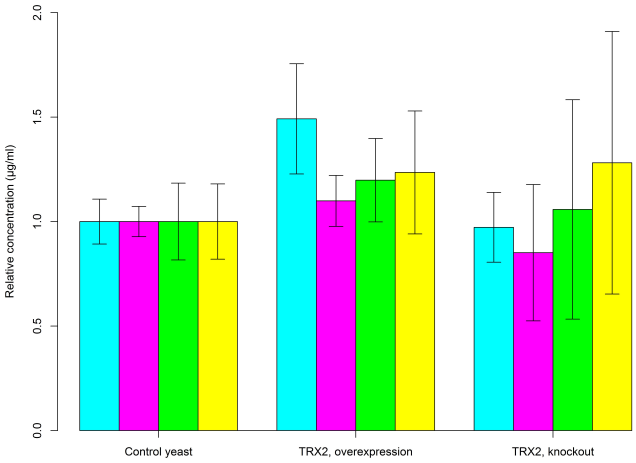

Figure 30: The relative antibody concentration of both the overexpression and knockout *TRX2* yeasts compared to control yeasts on each respective plate. Antibody secretion relative to control yeasts was higher on the overexpression yeasts across all galactose concentrations. The knockouts also had larger relative concentration of antibody on 3% and 4% galactose, but the error margins are high.

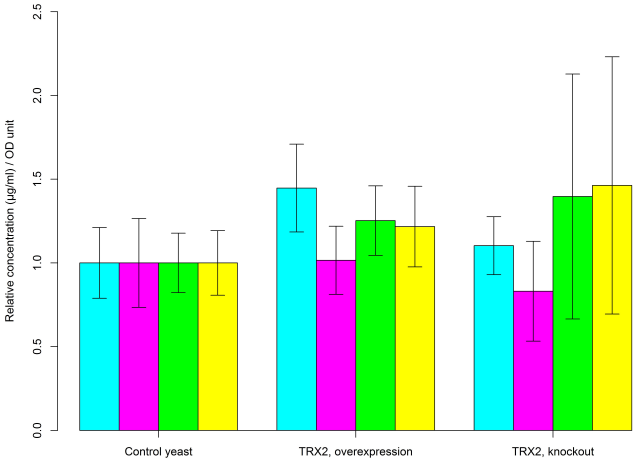

Figure 31: The relative antibody concentration per OD unit of both the overexpression and knockout *TRX2* yeasts compared to control yeasts on each respective plate. The overexpression of *TRX2* increased the antibody concentration per OD unit relative to control yeast. Also the knockout yeasts had increased antibody concentration per OD unit on 3% and 4% galactose concentrations, but the error margins are very large.

None of the genes selected for experiments *in vivo* had been tested in *S. cerevisiae* regarding antibody production. It was found in a study by Toikkanen et al. (2003)

that overexpressing *SBH1* enhanced the production of α-amylase in *S. cerevisiae*. In this study, a contradicting result was obtained, however as stated previously the effects of modifications are specific to both the host and target protein. In a study by Delic et al. (2014b) it was found that the overexpression of the gene encoding transcription factor Yap1, which regulates the expression of both *TRX1* and *TRX2*, improved protein secretion in the yeast *Pichia pastoris*. No studies regarding the effects of *GPD2* on protein production in yeast were found.

8 Conclusions and suggestions for future work

There are not yet any simple solutions on how to improve the production of proteins in cell factories. Different host species and different target proteins require different tactics for improving the product yield, as the bottleneck differs for each case. In this thesis naïve B cells were compared with fully matured plasma cells in the enrichment analysis in order to gain new ways to enhance the antibody production yield in *S. cerevisiae*.

The bottleneck in the production of IgG antibody in *S. cerevisiae* seems to lie in protein folding, as the overexpression of *TRX2* enhanced the antibody secretion. Further research could be aimed into the overexpression of other molecular chaperones involved in folding activities. In the enrichment analysis there were 24 genes found which had a homologue in *S. cerevisiae*. The 20 genes which were not studied *in vivo* in this study could be overexpressed and knocked out in yeast. Additionally, the mammalian genes which do not have a yeast homologue could be considered to be expressed in yeast. Also other enriched terms than the ones chosen here could be searched in order to find suitable candidates for genes to be expressed in yeast.

The results of the *in vivo* experiments showed that out of the genes tested, the overexpression of *TRX2* enhanced most the secretion of antibody in *S. cerevisiae*. Thioredoxins are involved in the folding of antibodies, therefore the bottleneck in the secretory pathway seems to be in the folding step. Also the results show that out of the two thioredoxins tested, as it enhances the antibody secretion *TRX2* seems to be more involved in the folding of antibodies than the more auxiliary *TRX1*, either by maintaining the redox homeostasis or by facilitating protein folding. As the knockout of *GPD2* improved IgG secretion, it could be combined with the overexpression of *TRX2* to see if further improved yield could be obtained. There was a lot of variation in antibody concentration between different rounds of ELISA, possibly due to freezing of the yeast strains. Even with the measures taken in this thesis to reduce the variation, the error margins in the results are still high. This is still preliminary data, and further reproduction of these experiments should reduce the variation. Also, as the successful transformation of the knockouts were verified only by growing on selective media, it is not completely sure if the transformations actually succeeded.

This thesis has shown that enrichment analysis provides useful information for improving the protein yield of cell factories. However, as the processes inside the cells are somewhat specific to species and in higher eukaryotes the type of cell, it is not always possible to directly apply the results of enrichment analysis into other type of cells. As this thesis has demonstrated, attempts to enhance the product yield are still trial and error based. This could be overcome by collecting a database from different studies, from which a researcher could search for suggested solutions for the unique combination of produced protein and the host species. Most of the enriched terms of plasma cells were involved in immunity, and not directly applicable for transformations into *S. cerevisiae*. Additional information could be gained from enrichment analysis comparing naïve B cells with developing plasma cells in different stages of development. As the ways to enhance protein secretion are dependent of species, cells of both wild type and antibody-secreting *S. cerevisiae* could be compared

with each other, and overexpress the enriched genes of the antibody-secreting yeast. Besides looking at transcriptomic data, enrichment analysis on proteomic data could be conducted, and those results could be tested *in vivo*.

References

Abbas, a. R., D. Baldwin, Y. Ma, W. Ouyang, a. Gurney, F. Martin, S. Fong, M. van Lookeren Campagne, P. Godowski, P. M. Williams, a. C. Chan, and H. F. Clark
2005. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes and immunity*, 6(4):319–31.

Amano, K., Y. Chiba, Y. Kasahara, Y. Kato, M. K. Kaneko, A. Kuno, H. Ito, K. Kobayashi, J. Hirabayashi, Y. Jigami, and H. Narimatsu
2008. Engineering of mucin-type human glycoproteins in yeast cells. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9):3232–7.

Anelli, T. and R. Sitia
2008. Protein quality control in the early secretory pathway. *The EMBO journal*, 27(2):315–27.

Ashburner, M., C. Ball, J. Blake, and D. Botstein
2000. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Berg, J., J. Tymoczko, and L. Stryer
2007. *Biochemistry, 6th edition.* New York: W.H. Freeman and Company.

Berlec, A. and B. Strukelj
2013. Current state and recent advances in biopharmaceutical production in Escherichia coli, yeasts and mammalian cells. *Journal of industrial microbiology & biotechnology*, 40(3-4):257–74.

Berndt, C., C. H. Lillig, and A. Holmgren
2008. Thioredoxins and glutaredoxins as facilitators of protein folding. *Biochimica et biophysica acta*, 1783(4):641–50.

Bolstad, B.
2004. *Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization.* PhD thesis, University of California, Berkeley.

Brewer, J. W. and L. M. Hendershot
2005. Building an antibody factory: a job for the unfolded protein response. *Nature immunology*, 6(1):23–9.

Carlson, M.
. *hgu133a2.db: Affymetrix Human Genome U133A 2.0 Array annotation data (chip hgu133a2).* R package version 2.14.0.

Carlson, M.
. *hgu133b.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133b).* R package version 2.14.0.

Chiba, Y. and H. Akeboshi
2009. Glycan engineering and production of 'humanized' glycoprotein in yeast cells. *Biological and Pharmaceutical Bulletin*, 32(5):786–795.

Chiba, Y., M. Suzuki, S. Yoshida, A. Yoshida, H. Ikenaga, M. Takeuchi, Y. Jigami, and E. Ichishima
1998. Production of Human Compatible High Mannose-type (Man5GlcNAc2) Sugar Chains in Saccharomyces cerevisiae. *Journal of Biological Chemistry*, 273(41):26298–26304.

Chigira, Y., T. Oka, T. Okajima, and Y. Jigami
2008. Engineering of a mammalian O-glycosylation pathway in the yeast Saccharomyces cerevisiae: production of O-fucosylated epidermal growth factor domains. *Glycobiology*, 18(4):303–14.

Cocco, M., S. Stephenson, M. a. Care, D. Newton, N. a. Barnes, A. Davison, A. Rawstron, D. R. Westhead, G. M. Doody, and R. M. Tooze
2012. In vitro generation of long-lived human plasma cells. *The Journal of Immunology*, 189(12):5773–5785.

Consortium, T. U.
2014. Activities at the Universal Protein Resource (UniProt). *Nucleic acids research*, 42(Database issue):D191–8.

Covens, K., B. Verbinnen, N. Geukens, I. Meyts, F. Schuit, L. Van Lommel, M. Jacquemin, and X. Bossuyt
2013. Characterization of proposed human B-1 cells reveals pre-plasmablast phenotype. *Blood*, 121(26):5176–83.

Cronwright, G. R., J. M. Rohwer, a. Bernard, and B. a. Prior
2002. Metabolic Control Analysis of Glycerol Synthesis in Saccharomyces cerevisiae Metabolic Control Analysis of Glycerol Synthesis in Saccharomyces cerevisiae. *Applied and Environmental Microbiology*, 68(9):4448–4456.

Davis, S. and P. S. Meltzer
2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics (Oxford, England)*, 23(14):1846–7.

de Ruijter, J. C. and A. D. Frey
2015. Analysis of antibody production in Saccharomyces cerevisiae: effects of ER protein quality control disruption. *Applied Microbiology and Biotechnology*.

Delic, M., R. Göngrich, D. Mattanovich, and B. Gasser
2014a. Engineering of protein folding and secretion-strategies to overcome bottlenecks for efficient production of recombinant proteins. *Antioxidants & redox signaling*, 21(3):414–37.

Delic, M., A. B. Graf, G. Koellensperger, and C. Haberhauer-troyer
2014b. Overexpression of the transcription factor Yap1 modifies intracellular redox conditions and enhances recombinant protein secretion. *Microbial Cell*, 1(11):376–386.

Dunning, M., A. Lynch, and M. Eldridge
. *illuminaHumanv4.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4)*. R package version 1.22.1.

Edgar, R., M. Domrachev, and A. E. Lash
2002. Gene Expression Omnibus : NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

Falcon, S. and R. Gentleman
2007. Using GOstats to test gene lists for GO term association. *Bioinformatics (Oxford, England)*, 23(2):257–8.

Falcon, S. and R. Gentleman
2008. In *Bioconductor Case Studies*, R. Gentleman, K. Hornik, and G. Parmigiani, eds., chapter Hypergeometric Testing Used for Gene Set Enrichment Analysis. New York: Springer.

Feige, M. J., L. M. Hendershot, and J. Buchner
2010. How antibodies fold. *Trends in biochemical sciences*, 35(4):189–98.

Finke, K., K. Plathl, S. Panznerl, S. Prehn, T. A. Rapoport, E. Hartmann, and T. Sommer
1996. A second trimeric complex containing homologs of the Sec6lp complex functions in protein transport across the ER membrane of S . cerevisiae. *The EMBO journal*, 15(7):1482–1494.

Frenken, L. G. J., R. H. J. V. D. Linden, P. W. J. J. Hermans, J. W. Bos, R. C. Ruuls, B. D. Geus, and C. T. Verrips
2000. Isolation of antigen specific Llama V HH antibody fragments and their high level secretion by Saccharomyces cere6isiae. *Journal of Biotechnology*, 78:11 – 21.

Garrido, E. O. and C. M. Grant
2002. Role of thioredoxins in the response of Saccharomyces cerevisiae to oxidative stress induced by hydroperoxides. *Molecular Microbiology*, 43(4):993–1003.

Gasser, B., M. Saloheimo, U. Rinas, M. Dragosits, E. Rodríguez-Carmona, K. Baumann, M. Giuliani, E. Parrilli, P. Branduardi, C. Lang, D. Porro, P. Ferrer, M. L. Tutino, D. Mattanovich, and A. Villaverde
2008. Protein folding and conformational stress in microbial cells producing recombinant proteins: a host comparative overview. *Microbial cell factories*, 7:11.

Gentleman, R., V. Carey, W. Huber, and F. Hahne
. *genefilter: genefilter: methods for filtering genes from microarray experiments*. R package version 1.46.1.

Ghazalpour, A., B. Bennett, V. A. Petyuk, L. Orozco, R. Hagopian, I. N. Mungrue, C. R. Farber, J. Sinsheimer, H. M. Kang, N. Furlotte, C. C. Park, P.-Z. Wen, H. Brewer, K. Weitz, D. G. Camp, C. Pan, R. Yordanova, I. Neuhaus, C. Tilford, N. Siemers, P. Gargalovic, E. Eskin, T. Kirchgessner, D. J. Smith, R. D. Smith, and A. J. Lusis
2011. Comparative Analysis of Proteome and Transcriptome Variation in Mouse. *PLoS Genetics*, 7(6):e1001393.

Gitlin, A. D., Z. Shulman, and M. C. Nussenzweig
2014. Clonal selection in the germinal centre by regulated proliferation and hypermutation. *Nature*, 509(7502):637–40.

Good, K. L., D. T. Avery, and S. G. Tangye
2009. Programmed for Enhanced Survival and Responsiveness to Diverse Stimuli Compared to Naive B Cells 1. *The Journal of Immunology*, 182(2):890–901.

Güldener, U., S. Heck, T. Fiedler, J. Beinhauer, J. H. Hegemann, J.-l.-u. Gießen, and F. Straße
1996. A new efficient gene disruption cassette for repeated use in budding yeast. *Nucleic acids research*, 24(13):2519–2524.

Gygi, S. P., Y. Rochon, B. R. Franza, and R. Aebersold
1999. Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology*, 19(3):1720–1730.

Hackel, B. J., D. Huang, J. C. Bubolz, X. X. Wang, and E. V. Shusta
2006. Production of soluble and active transferrin receptor-targeting single-chain antibody using Saccharomyces cerevisiae. *Pharmaceutical Research*, 23(4):790–797.

Haider, S. and R. Pal
2013. Integrated analysis of transcriptomic and proteomic data. *Current genomics*, 14(2):91–110.

Hetz, C.
2012. The unfolded protein response: controlling cell fate decisions under ER stress and beyond. *Nature Reviews Molecular Cell Biology*, 13(2):89–102.

Hibi, T. and H.-M. Dosch
1986. Limiting dilution analysis of the b cell compartment in human bone marrow. *European Journal of Immunology*, 16(2):139–145.

Horgan, R. P. and L. C. Kenny
2011. 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, 13(3):189–195.

Hou, J., K. E. J. Tyo, Z. Liu, D. Petranovic, and J. Nielsen
2012. Metabolic engineering of recombinant protein secretion by Saccharomyces cerevisiae. *FEMS yeast research*, 12(5):491–510.

Huang, D. W., B. T. Sherman, and R. a. Lempicki
2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13.

Hubmann, G., S. Guillouet, and E. Nevoigt
2011. Gpd1 and Gpd2 fine-tuning for sustainable reduction of glycerol formation in Saccharomyces cerevisiae. *Applied and environmental microbiology*, 77(17):5857–67.

Idiris, A., H. Tohda, H. Kumagai, and K. Takegawa
2010. Engineering of protein secretion in yeast: strategies and impact on protein production. *Applied microbiology and biotechnology*, 86(2):403–17.

Janeway, C., P. Travers, M. Walport, and M. Shlomchik
2001. *Immunobiology: The Immune System in Health and Disease*. New York: Garland Science.

Jønson, L., J. F. Rehfeld, and A. H. Johnsen
2004. Enhanced peptide secretion by gene disruption of CYM1, a novel protease in Saccharomyces cerevisiae. *European journal of biochemistry / FEBS*, 271(23-24):4788–97.

Kang, H., E. Choi, W.-K. Hong, K. J.-Y., S.-M. Ko, J.-H. Sohn, and S. K. Rhee
2000. Proteolytic stability of recombinant human serum albumin secreted in the yeast Saccharomyces cerevisiae. *Applied microbiology and biotechnology*, 53(5):575–582.

Klein, U., S. Casola, G. Cattoretti, Q. Shen, M. Lia, T. Mo, T. Ludwig, K. Rajewsky, and R. Dalla-Favera
2006. Transcription factor IRF4 controls plasma cell differentiation and class-switch recombination. *Nature immunology*, 7(7):773–82.

Koskela, E. V. and A. D. Frey
2014. Homologous Recombinatorial Cloning Without the Creation of Single-Stranded Ends : Exonuclease and Ligation-Independent Cloning (ELIC).

Lin, K.-I., C. Angelin-Duclos, T. C. Kuo, and K. Calame
2002. Blimp-1-Dependent Repression of Pax-5 Is Required for Differentiation of B Cells to Immunoglobulin M-Secreting Plasma Cells. *Molecular and Cellular Biology*, 22(13):4771–4780.

MacDonald, J. W.
. *hugene10sttranscriptcluster.db: Affymetrix hugene10 annotation data (chip hugene10sttranscriptcluster)*. R package version 8.1.0.

Madigan, M. T., J. M. Martinko, P. V. Dunlap, and D. P. Clark
2009. *Brock Biology of Microorganisms 12th International Edition*. San Francisco: Pearson Education, Inc.

Malone, J. H. and B. Oliver
2011. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, 9(34).

Masciarelli, S. and R. Sitia
2008. Building and operating an antibody factory: redox control during B to plasma cell terminal differentiation. *Biochimica et biophysica acta*, 1783(4):578–88.

Mumberg, D., R. Müller, and M. Funk
1995. Yeast vectors for controlled expression of heterologous proteins in different genetic backgrounds. *Gene*, 156(1):119–122.

Nakanishi-shindo, Y., K.-i. Nakayamas, A. Tanakag, and Y. Todag
1993. Structure of the N-Linked Oligosaccharides That Show the Complete Loss of alpha-1,6-Polymannose Outer Chain from och1, och1 mnn1, and och1 mnn1 alg3 Mutants of Saccharomyces cerevisiae. *The Journal of biological chemistry*, 268(35):26338–26345.

Nature Publishing Group
2014. Nature.com subject areas: Transcriptomics.

Ng, D. T. W., J. D. Brown, and P. Walter
1996. Signal Sequences Specify the Targeting Route to the Endoplasmic Reticulum Membrane. *The Journal of Cell Biology*, 134(2):269–278.

Nutt, S. L., N. Taubenheim, J. Hasbold, L. M. Corcoran, and P. D. Hodgkin
2011. The genetic network controlling plasma cell differentiation. *Seminars in immunology*, 23(5):341–9.

Oracki, S. a., J. a. Walker, M. L. Hibbs, L. M. Corcoran, and D. M. Tarlinton
2010. Plasma cell development and survival. *Immunological reviews*, 237(1):140–59.

Parsaie Nasab, F., M. Aebi, G. Bernhard, and A. D. Frey
2013. A combined system for engineering glycosylation efficiency and glycan structure in Saccharomyces cerevisiae. *Applied and environmental microbiology*, 79(3):997–1007.

Piirainen, M. A., J. C. de Ruijter, E. V. Koskela, and A. D. Frey
2014. Glycoengineering of yeasts from the perspective of glycosylation efficiency. *New Biotechnology*, 31(6):532–537.

Quackenbush, J.
2006. Computational approaches to analysis of DNA microarray data. *Yearbook of medical informatics*, Pp. 91–103.

R Core Team
2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reimold, A., N. Iwakoshi, J. Manis, P. Vallabhajosyula, E. Szomolanyi-Tsuda, E. M. Gravallese, D. Friend, M. J. Grusby, F. Alt, and L. H. Glimcher
2001. Plasma cell differentiation requires the transcription factor XBP-1. *Nature*, 412(6844):300–307.

Romijn, E. P., C. Christis, M. Wieffer, J. W. Gouw, A. Fullaondo, P. van der Sluijs, I. Braakman, and A. J. R. Heck
2005. Expression clustering reveals detailed co-expression patterns of functionally related proteins during B cell differentiation: a proteomic study using a combination of one-dimensional gel electrophoresis, LC-MS/MS, and stable isotope labeling by amino acid. *Molecular & cellular proteomics : MCP*, 4(9):1297–310.

Rose, M. D., L. M. Misra, and J. P. Vogel
1989. KAR2, a karyogamy gene, is the yeast homolog of the mammalian BiP/GRP78 gene. *Cell*, 57(7):1211–1221.

Sciammas, R., a. L. Shaffer, J. H. Schatz, H. Zhao, L. M. Staudt, and H. Singh
2006. Graded expression of interferon regulatory factor-4 coordinates isotype switching with plasma cell differentiation. *Immunity*, 25(2):225–36.

Shaffer, A., K.-I. Lin, T. C. Kuo, X. Yu, E. M. Hurt, A. Rosenwald, J. M. Giltnane, L. Yang, H. Zhao, K. Calame, and L. M. Staudt
2002. Blimp-1 Orchestrates Plasma Cell Differentiation by Extinguishing the Mature B Cell Gene Expression Program. *Immunity*, 17(1):51–62.

Shaffer, A. L., M. Shapiro-Shelef, N. N. Iwakoshi, A.-H. Lee, S.-B. Qian, H. Zhao, X. Yu, L. Yang, B. K. Tan, A. Rosenwald, E. M. Hurt, E. Petroulakis, N. Sonenberg, J. W. Yewdell, K. Calame, L. H. Glimcher, and L. M. Staudt
2004. XBP1, downstream of Blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation. *Immunity*, 21(1):81–93.

Shapiro-Shelef, M., K.-I. Lin, L. J. McHeyzer-Williams, J. Liao, M. G. McHeyzer-Williams, and K. Calame
2003. Blimp-1 Is Required for the Formation of Immunoglobulin Secreting Plasma Cells and Pre-Plasma Memory B Cells. *Immunity*, 19(4):607–620.

Shusta, E. V., R. T. Raines, A. Plückthun, and K. D. Wittrup
1998. Increasing the secretory capacity of Saccharomyces cerevisiae for production of single-chain antibody fragments. *Nature Biotechnology*, 16:773–777.

Spadiut, O., S. Capone, F. Krainer, A. Glieder, and C. Herwig
2014. Microbials for the production of monoclonal antibodies and antibody fragments. *Trends in biotechnology*, 32(1):54–60.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, and B. L. Ebert
2005. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550.

Taxis, C. and M. Knop
2006. System of centromeric, episomal, and integrative vectors based on drug resistance markers for Saccharomyces cerevisiae. *BioTechniques*, 40(1):73–78.

Thomassen, Y. E., W. Meijer, L. Sierkstra, and C. T. Verrips
2002. Large-scale production of VHH antibody fragments by Saccharomyces cerevisiae. *Enzyme and Microbial Technology*, 30(3):273–278.

Toikkanen, J., E. Gatti, K. Takei, M. Saloheimo, V. M. Olkkonen, H. Söderlund, P. De Camilli, and S. Keränen
1996. Yeast protein translocation complex: isolation of two genes SEB1 and SEB2 encoding proteins homologous to the Sec61 beta subunit. *Yeast (Chichester, England)*, 12(5):425–38.

Toikkanen, J. H., K. J. Miller, H. Söderlund, J. Jäntti, and S. Keränen
2003. The beta subunit of the Sec61p endoplasmic reticulum translocon interacts with the exocyst complex in Saccharomyces cerevisiae. *The Journal of biological chemistry*, 278(23):20946–53.

Tortora, G. and B. Derrickson
2006. *Principles of Anatomy and Physiology.* Hoboken: John Wiley and Sons, Inc.

Trotter, E. W. and C. M. Grant
2002. Thioredoxins are required for protection against a reductive stress in the yeast Saccharomyces cerevisiae. *Molecular Microbiology*, 46(3):869–878.

van Anken, E., E. P. Romijn, C. Maggioni, A. Mezghrani, R. Sitia, I. Braakman, and A. J. Heck
2003. Sequential Waves of Functionally Related Proteins Are Expressed When B Cells Prepare for Antibody Secretion. *Immunity*, 18(2):243–253.

Vemuri, G. N., M. A. Eiteman, J. E. McEwen, L. Olsson, and J. Nielsen
2007. Increasing NADH oxidation reduces overflow metabolism in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 104(7):2402–7.

Victora, G. D. and M. C. Nussenzweig
2012. Germinal centers. *Annual review of immunology*, 30:429–57.

Walsh, G.
2010. Post-translational modifications of protein biopharmaceuticals. *Drug Discovery Today*, 15(17-18):773–780.

Walsh, G.
2014. Biopharmaceutical_benchmarks_2010. *Nature Biotechnology*, 32(10):992–1000.

Wang, Z., M. Gerstein, and M. Snyder
2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

Xu, P., D. Raden, F. J. Doyle, and A. S. Robinson
2005. Analysis of unfolded protein response during single-chain antibody expression in Saccaromyces cerevisiae reveals different roles for BiP and PDI in folding. *Metabolic engineering*, 7(4):269–79.

Yoshida, H., T. Matsui, A. Yamamoto, T. Okada, and K. Mori
2001. XBP1 mRNA Is Induced by ATF6 and Spliced by IRE1 in Response to ER Stress to Produce a Highly Active Transcription Factor. *Cell*, 107(7):881–891.

Young, C. L. and A. S. Robinson
2014. Protein folding and secretion: mechanistic insights advancing recombinant protein production in S. cerevisiae. *Current opinion in biotechnology*, 30C:168–177.

Zhang, X., N. Bjorklund, G. Alvare, T. Ryzdak, R. Sparling, and B. Fristensky
2014. *ELBOW: ELBOW - Evaluating foLd change By the lOgit Way*. R package version 1.0.0.

A Gene enrichment analysis

```
#1 LOADING REQUIRED PACKAGES

install.packages("survival")

#this package is for downloading the data from GEO forming
#an ExpressionSet-object
source("http://bioconductor.org/biocLite.R")
biocLite("GEOquery")
require(GEOquery)
library(GEOquery)

#this package includes function that allow normalization of
#an ExpressionSet-object
biocLite("affyPLM")
require(affyPLM)
library(affyPLM)

#the annotation data
biocLite("illuminaHumanv4.db")
require(illuminaHumanv4.db)
biocLite("hugene10sttranscriptcluster.db")
library(hugene10sttranscriptcluster.db)
biocLite("hgu133a2.db")
library(hgu133a2.db)
biocLite("hgu133b.db")
library(hgu133b.db)

#this package is for enrichment analysis
biocLite("GOstats")
require(GOstats)
library(GOstats)

#fold change analysis package
biocLite("ELBOW")
library("ELBOW")

#for t-tests
require(genefilter)

biocLite("Rgraphviz")

#2 ANALYSIS OF DATA
```

```
#Downloading data sets to R
#Cocco's data
gset <- getGEO(filename= "GSE41208_series_matrix.txt", GSEMatrix =TRUE)
gset@annotation <- "illuminaHumanv4.db"

#filtering of data: removing probes without Entrez-ID
entrezIds <- mget(featureNames(gset), envir=illuminaHumanv4ENTREZID)
haveEntrezId <- names(entrezIds)[sapply(entrezIds, function(x) !is.na(x))]
gset.fil <- gset[haveEntrezId, ]

gset.fil@phenoData$STATUS #check if this is null, so can be utilized
#Mark the different experimental points
gstatus <- c(1,1,1,3,3,3,6,6,6,13,13,13,20,20,41,41,41,42,42,42,42)
gset.fil@phenoData$STATUS <- gstatus
#The expression data are marked by experimental point

CC1 <-gset.fil[, gset.fil@phenoData$STATUS%in%c(1,20)]
#comparing time point 0 with time point 20 d
CC2 <-gset.fil[, gset.fil@phenoData$STATUS%in%c(1,41)]
#comparing time point 0 with time point 41 d
CC3 <-gset.fil[, gset.fil@phenoData$STATUS%in%c(1,42)]
#comparing time point 0 with bone marrow plasma cells

#GO-enrichment analysis

#CC1

#Finding differentially expressed genes for GO-analysis
my.dataCC1 = exprs(CC1)
my.classes = c(1,1,1,20,20)
genenames = rownames(my.dataCC1)
ttestCutoff <- 0.0003
ttests = rowttests(my.dataCC1, factor(my.classes))
smPV = ttests$p.value < ttestCutoff
pvalFilteredCC1 <- my.dataCC1[smPV, ]

#Formatting data for fold change analysis
write.table(pvalFilteredCC1, file="ttestgenesCC1BP.csv", row.names=TRUE,
            append=FALSE, sep=",")
csv_data <- read.table("ttestgenesCC1BP.csv", header=TRUE, sep=",", dec=".",
                       row.names=NULL)

init_count <- 3 #number of initial condition columns
final_count <- 2 #number of final condition columns
working_sets <- extract_working_sets(csv_data, init_count, final_count)
```

```
probes <- working_sets[[1]]
initial_conditions <- working_sets[[2]]
final_conditions <- working_sets[[3]]

#Finding genes with significant fold change
sig <- analyze_elbow(probes, initial_conditions, final_conditions)
write.table(sig,file="signprobesCC1BP.csv",sep=",",row.names=FALSE,
            append=FALSE)
write.table(sig$probe_names,file="signprobenamesCC1BP.csv",sep=",",
            row.names=FALSE, col.names=FALSE, quote=FALSE, append=FALSE)
selectedgenesCC1BP <- readLines("signprobenamesCC1BP.csv")

selectedEntrezIds <- unlist(mget(selectedgenesCC1BP,illuminaHumanv4ENTREZID))
entrezUniverse <- unlist(mget(genenames,illuminaHumanv4ENTREZID))

#Parameters are set up for GO-enrichment testing
hgCutoff <- 0.00001
params <- new("GOHyperGParams", geneIds=selectedEntrezIds,
              universeGeneIds=entrezUniverse,
              annotation="illuminaHumanv4.db",
              ontology="BP",
              pvalueCutoff=hgCutoff,
              conditional=TRUE,
              testDirection="over")
#gives warning because of duplicates: removes them
#Enrichment test
hgOverCC1BP <- hyperGTest(params)
hgOverCC1BP
termListCC1BP <- summary(hgOverCC1BP)
termListCC1BP
htmlReport(hgOverCC1BP, file="hgOverCC1BP.html")

#two categories are tested for each differentiation state
#Finding differentially expressed genes for GO-analysis
ttestCutoff <- 0.001
smPV = ttests$p.value < ttestCutoff
pvalFilteredCC1 <- my.dataCC1[smPV, ]

#Formatting data for fold change analysis
write.table(pvalFilteredCC1, file="ttestgenesCC1MF.csv", row.names=TRUE,
            append=FALSE, sep=",")
csv_data <- read.table("ttestgenesCC1MF.csv", header=TRUE, sep=",", dec=".",
                       row.names=NULL)
working_sets <- extract_working_sets(csv_data, init_count, final_count)
probes <- working_sets[[1]]
```

```
initial_conditions <- working_sets[[2]]
final_conditions <- working_sets[[3]]

#Finding genes with significant fold change
sig <- analyze_elbow(probes, initial_conditions, final_conditions)
write.table(sig,file="signprobesCC1MF.csv",sep=",",row.names=FALSE,
            append=FALSE)
write.table(sig$probe_names,file="signprobenamesCC1MF.csv", append=FALSE,
            sep=",",row.names=FALSE, col.names=FALSE, quote=FALSE)
selectedgenesCC1MF <- readLines("signprobenamesCC1MF.csv")

selectedEntrezIds <- unlist(mget(selectedgenesCC1MF,illuminaHumanv4ENTREZID))

#Parameters are set up for GO-enrichment testing
hgCutoffMF <- 0.001
params <- new("GOHyperGParams", geneIds=selectedEntrezIds,
                universeGeneIds=entrezUniverse,
                annotation="illuminaHumanv4.db",
                ontology="MF",
                pvalueCutoff=hgCutoffMF,
                conditional=TRUE,
                testDirection="over")

#Enrichment test
hgOverCC1MF <- hyperGTest(params)
hgOverCC1MF
termListCC1MF <- summary(hgOverCC1MF)
termListCC1MF
htmlReport(hgOverCC1MF, file="hgOverCC1MF.html")

#Repeat the process for other data subsets
```

B Differential expression

```
#LISTS OF DIFFERENTIALLY EXPRESSED GENES WITH THE LOWEST P-VALUES

#This function determines the differentially expressed genes with the lowest
#P-values.
#Inputs:
#data = gene expression data
#classes = experimental points
#ttestCutoff = cut-off value for rowttests function
#ann = annotation of data, as character, e.g. "illuminaHumanv4"
lowestPvalues <- function(data, classes, ttestCutoff, ann) {
  ttests = rowttests(data, factor(classes))
  smPV = ttests$p.value < ttestCutoff
  pvalFiltered <- data[smPV, ]
  selectedgenes = rownames(pvalFiltered)
  namelist <- eval(parse(text=paste(c("mget(selectedgenes,", ann, "GENENAME)"),
                                    collapse="")))
  result <- as.matrix(namelist)
  colnames(result, do.NULL = TRUE)
  colnames(result) <- "Gene"
  result
}

ttestCutoff <- 0.000002

my.classes = c(1,1,1,20,20)
namelistCC1 <- lowestPvalues(data = my.dataCC1, classes = my.classes,
                             ttestCutoff = ttestCutoff, ann = "illuminaHumanv4")

my.classes = c(1,1,1,41,41,41)
namelistCC2 <- lowestPvalues(data = my.dataCC2, classes = my.classes,
                             ttestCutoff = ttestCutoff, ann = "illuminaHumanv4")

my.classes = c(1,1,1,42,42,42,42)
namelistCC3 <- lowestPvalues(data = my.dataCC3, classes = my.classes,
                             ttestCutoff = ttestCutoff, ann = "illuminaHumanv4")

my.classes = c(1,1,1,4,4,4)
namelistC1 <- lowestPvalues(data = my.dataC1, classes = my.classes,
                            ttestCutoff = ttestCutoff,
                            ann = "hugene10sttranscriptcluster")

ttestCutoff <- 0.0001
my.classes = c(1,1,4,4)
```

```
namelistGT1 <- lowestPvalues(data = my.dataGT1, classes = my.classes,
                             ttestCutoff = ttestCutoff, ann = "hgu133a2")

ttestCutoff <- 0.000002
my.classes = c(rep((2),7),rep((4),3))
namelistA1 <- lowestPvalues(data = my.dataA1, classes = my.classes,
                            ttestCutoff = ttestCutoff, ann = "hgu133a2")

my.classes = c(rep((2),7),rep((5),4))
namelistA2 <- lowestPvalues(data = my.dataA2, classes = my.classes,
                            ttestCutoff = ttestCutoff, ann = "hgu133a2")

my.classes = c(rep((2),7),rep((4),3))
namelistA3 <- lowestPvalues(data = my.dataA3, classes = my.classes,
                            ttestCutoff = ttestCutoff, ann = "hgu133b")

my.classes = c(rep((2),7),rep((5),4))
namelistA4 <- lowestPvalues(data = my.dataA4, classes = my.classes,
                            ttestCutoff = ttestCutoff, ann = "hgu133b")
```

C termLists function

```
#This function compares the enriched GO-terms of each data subset
#Returns a table, where there is one GO-term in each row, and the columns
#show in which subsets the term appears in. Terms are arranged according to
#the number of enriched term lists they appear in.
#Inputs:
#tlist = list that contains sublists of enriched terms at different data subsets
#names = names of the data subsets in the order in which they appear in tlist


termLists <- function(tlist, names){
  #An empty matrix for saving the results
  result <- matrix(data = NA, nrow = 1000, ncol = length(tlist)+2)
  dimnames(result) <- list(c(NULL), c("Genes", "Count",
                                      paste("Termlist", 1:length(tlist),
                                            sep="")))

  for(i in 1:nrow(result)){ #Sets up a counter
    result[i,2] = 0
    i = i + 1
  }


  for(i in 1:length(tlist)) { #Goes through the gene lists one by one
    x <- tlist[i] #Current gene list
    terms <- tlist[[i]]$Term #Terms that appear in current gene list
    id <- names[[i]] #The name of current gene list

    for(j in 1:length(terms)) { #Goes through the term list one by one
      for(k in 1:dim(result)[1]) {#Compares the terms to those in result matrix
        matchFound = FALSE #This checks whether a match was found on the list
        comparison <- terms[j] == result[k,1]
        if(is.na(comparison)==TRUE) {
          comparison=FALSE
        }
        if(comparison == TRUE){ #Checks for a match
          matchFound = TRUE
          for(m in 3:dim(result)[2]) {
            if(is.na(result[k,m])==TRUE) {
              #Saves the gene list id to the first available column
              result[k,m]=id
              break
            } else {
              m = m+1
```

```
          }
        }
        break
      } else {
        #If no match is found, checks the next term in the result matrix
        k= k+1
      }
    }
    if(matchFound==FALSE){
      #If no matches were found, adds the term and gene list id to
      #the first available spot in the result matrix
      for(n in 1:dim(result)[1]) {
        if(is.na(result[n,1])==TRUE) {
          result[n,1]=terms[j]
          result[n,3]=id
          break
        } else {
          n = n+1
        }
      }
    }
    j = j+1 #Moves on to the next term
  }
  i = i+1 #Moves on to the next list
}

#Omit the NA rows from result matrix
matrixNA <- complete.cases(result[ ,1])
result2 <- result[matrixNA, ]

#Omit the NA columns from result matrix
countNA = 0
for(j in 1:ncol(result2)){
  amountNA = sum(!is.na(result2[ ,j]))
  empty <- amountNA == 0
  if(empty == TRUE){
    countNA = countNA + 1
    j = j+1
  } else {
    j = j+1
  }
}
firstDeletedColumn = ncol(result2)-countNA+1
result2 <- result2[,-c(firstDeletedColumn:ncol(result2))]
```

```
  #Counts the amount of lists in which each term appears
  for(i in 1:nrow(result2)){
    counter = sum(!is.na(result2[i, ]))-2
    result2[i,2]=counter
    i = i+1
  }

  #Arranges the rows
  result4 <- transform.default(result2)
  order.count <- order(result4$Count, decreasing=TRUE)
  result5 <- result4[order.count, ]
}
```

D extractGenes function

```
#This function checks which genes comprise the selected GO-term.
#It returns a table that has one gene in each row, and columns show in which
#subsets the gene appears in. If the GO-term is not enriched in a subset,
#it is not counted in the table. Genes are arranged according to the number
#of subsets they appear in.
#Parameters:
#tlist = list of GOHyperGResult objects
#selgenes = lists of genes used in GO-enrichment testing
#names = names of data subsets
#GOid = the GO ID of the selected GO-term

extractGenes <- function(tlist, selgenes, names, GOid){
  lst <- list() #An empty list for saving the results

  #Goes through the GOHyperGResult objects one by one and extracts their
  #annotations
  for(i in 1:length(tlist)) {
    ann <- annotation(tlist[[i]])

    #This function lists all Probe Set IDs associated with the selected
    #Entrez IDs annotated at each significant GO term in the test result.
    probes <- probeSetSummary(tlist[[i]], sigProbesets = selgenes[[i]],
                              ids = "ENTREZID")
    selectedprobes <- probes[[GOid]] #All probes associated with GOid
    empty <- is.null(selectedprobes)

    if(all(empty) == FALSE) {
      dname <- names[[i]] #Name of the current subset

      #Extracts the names of the genes, creates a sublist for them
      probeIds = selectedprobes[ ,2]
      lst[[dname]] <- eval(parse(text=paste(c("mget(probeIds,", ann,
                                             "GENENAME)"), collapse="")))
      i = i+1 #Moves on to the next subset
    } else { #If true, GOid was not enriched in the current subset
        i = i+1
      }
  }

  lst[sapply(lst, is.null)] <- NULL  #Removes null elements
  result <- geneLists(lst) #Calls geneLists function to final result format
  result
}
```

E geneLists function

```
#This function calculates the amount on data subsets that a gene appears in.
#It returns a table containing genes, the amount of data subsets they appear
#in, and the corresponding subsets.
#Inputs
#glist = a list containing sublists of genes that appear in specific
#        data subsets

geneLists <- function(glist) {
  resultLength = 0
  for(i in 1:length(glist)){
    resultLength = resultLength + length(glist[[i]])
  }
  #An empty matrix for saving the results
  result <- matrix(data = NA, nrow = resultLength, ncol = 2+length(glist))
  dimnames(result) <- list(c(NULL), c("Genes", "Amount of subsets",
                                      paste("Genelist", 1:length(glist),
                                            sep="")))

  #Sets up the result matrix to contain the names of all genes in the lists
  k = 1
  for(i in 1:length(glist)) {
    for(j in 1:length(glist[[i]])) {
      gene <- as.character(glist[[i]][j])
      result[k,1] = gene
      k = k+1
      j = j+1
    }
    i = i+1
  }


  for(i in 1:nrow(result)) { #Goes through the result matrix
    for(j in 1:length(glist)) {
      id = names(glist)[[j]] #Name of the current data subset
      match = result[i,1]%in%glist[[j]]
      if(match == TRUE){
        j = j+1
        for(m in 3:dim(result)[2]) {
          if(is.na(result[i,m])==TRUE) {
            #Saves the data subset name to the first available column
             result[i,m]=id
             break
          } else {
```

```
            m = m+1
          }
        }
      } else {
        j = j+1 #Moves on to the next data subset
      }
    }
    i = i+1 #Moves on to the next gene
    }

result2 = unique(result) #Omits duplicate genes

#Counts the amount of data subsets in which each gene appears
  for(i in 1:nrow(result2)){
    counter = sum(!is.na(result2[i, ]))-1
    result2[i,2]=counter
    i = i+1
  }

#Omits the NA columns from result matrix
countNA = 0
 for(j in 1:ncol(result2)){
   amountNA = sum(!is.na(result2[ ,j]))
   empty <- amountNA == 0
     if(empty == TRUE){
     countNA = countNA + 1
     j = j+1
   } else {
     j = j+1
   }
 }
 firstDeletedColumn = ncol(result2)-countNA+1
 if(firstDeletedColumn > ncol(result2)) {
   result4 <- transform.default(result2)
 } else {
 result2 <- result2[,-c(firstDeletedColumn:ncol(result2))]
 result4 <- transform.default(result2)
 }

#Rearranges the rows so that in the first row is the gene that appears
#in the largest amount data subsets
order.count <- order(result4[ ,2], decreasing=TRUE)
result5 <- result4[order.count, ]
}
```

F compareGeneLists function

```
#This function compares the differentially expressed genes and the genes
#associated with a certain GO-term. It returns a matrix where on each row
#there is a gene name (all genes are associated to the same GO-term), and
#the columns show in how many lists of differentially expressed genes
#the gene appears, in and the corresponding names of data subsets.
#It also tells in how many data subsets the gene was enriched.
#Inputs:
#genematrix = table of genes and data subsets in which they appear in
#(e.g. table returned by extractGenes function)
#diffList = a list where each sublist contains the list of most differentially
#expressed genes of a data subset
#names = names of data subsets in the order they appear in diffList


compareGeneLists <- function(genematrix, diffList, names) {
  #An empty matrix for saving the results
  result <- matrix(data = NA, nrow = length(genematrix[ ,1]),
                   ncol = 3+length(diffList))
  dimnames(result) <- list(c(NULL), c("Genes", "Enrichment subsets",
                                      "Differential expression subsets",
                                      paste("Genelist", 1:length(diffList),
                                            sep="")))
  #Sets up the result matrix to contain the names of all genes associated
  #with the GO term
  for(i in 1:length(genematrix[ ,1])) {
      gene <- as.character(genematrix[ ,1][i])
      result[i,1] = gene
      result[i,2] = as.character(genematrix[ ,2][i])
      i = i+1
  }

  for(i in 1:nrow(result)) { #Goes through the result matrix
    for(j in 1:length(diffList)) {
      id = names[[j]] #Name of the list of differentially expressed genes
      match = result[i,1]%in%diffList[[j]]
      if(match == TRUE){
        j = j+1
        for(m in 4:dim(result)[2]) {
            if(is.na(result[i,m])==TRUE) {
              #Saves the gene list id to the first available column
              result[i,m]=id
              break
            } else {
```

```
            m = m+1
          }
        }
      } else {
        j = j+1
      }
  }
  i = i+1
}

#Counts the amount of lists in which each gene appears
for(i in 1:nrow(result)){
  counter = sum(!is.na(result[i, ]))-2
  result[i,3]=counter
  i = i+1
}

#Omit the NA columns from result matrix
countNA = 0
for(j in 1:ncol(result)){
  amountNotNA = sum(!is.na(result[ ,j]))
  empty <- amountNotNA == 0 #If a column is empty, number of non-NA rows is 0
  if(empty == TRUE){ #True if an empty column is found
    countNA = countNA + 1
    j = j+1
  } else {
    j = j+1
  }
}
firstDeletedColumn = ncol(result)-countNA+1
if(firstDeletedColumn > ncol(result)) {
  result <- transform.default(result)
} else {
  result <- result[,-c(firstDeletedColumn:ncol(result))]
  result2 <- transform.default(result)
}

#Rearranges the rows so that in the first row is the gene that appears
#in the largest amount of lists of differentially expressed genes
order.count <- order(result2[ ,3], decreasing=TRUE)
result3 <- result2[order.count, ]
}
```

G findFoldChange function

```
#This function finds the different fold changes of every gene in a list.
#It returns a table that has one gene in each row, and the fold changes
#in different data subsets in columns. It also calculates the mean and
#standard deviation of each gene's fold change.
#Inputs:
#glist = the list of genes which fold changes need to be found out
#foldChanges = lists of fold changes
#names =  list of names of the data subsets
#annotations = list of annotations of the data subsets as characters,
#e.g. "illuminaHumanv4"

findFoldChange <- function(glist, foldChanges, names, annotations) {

  #An empty matrix for saving the results
  result <- matrix(data = NA, nrow = length(glist[ ,1]),
                   ncol = 3+length(foldChanges))
  dimnames(result) <- list(c(NULL), c("Gene", "Mean fold change",
                                      "SD of fold change", names))

  #these are for calculating each genes' mean fold change and
  #standard deviation
  rowmean <- vector()
  meanlist <- list()

  #Sets up the result matrix to contain the names of all genes in
  #the original list
  for(i in 1:length(glist[ ,1])) {
    gene <- as.character(glist[ ,1][i])
    result[i,1] = gene
    meanlist[[gene]] <- rowmean
    i = i+1
  }

  for(j in 1:length(foldChanges)) {
      #First find out the gene names of current foldChanges object
      current <- read.table(file = foldChanges[[j]], header=TRUE, sep=",")
      probeNames = as.character(current[ ,1])
      #Extract the fold changes of current foldChanges object
      fcs = as.numeric(current[, 2])
      #the change is calculated from final to initial state, so
      #multiply by -1 to get change from initial to final state
      fcs = fcs*(-1)
      ann = annotations[[j]]
```

```
      genenames <- eval(parse(text=paste(c("mget(probeNames,", ann,
                                           "GENENAME)"), collapse="")))
      fc <- matrix(data = NA, nrow = length(genenames), ncol=2)
      #Matrix fc contains gene names and fold changes of
      #the current foldChanges object
      for(k in 1:length(genenames)){
        fc[k,1] = as.character(genenames[[k]])
        fc[k,2] = as.character(current[k,2])
        k=k+1
      }

      for(i in 1:nrow(result)) { #Goes through the result matrix
        #Finds out if there is a fold change for that gene
        match = fc[ ,1]%in%result[i,1]

        if(any(match) == TRUE){
          if(is.null(dim(fc[match, ]))==TRUE) { #Only one match is found
            fold <- fc[match, ][[2]]
            meanlist[[i]] <- c(meanlist[[i]], as.numeric(fc[match, ][[2]]))
          } else{ #Multiple matches
            fold <- paste(fc[match, ][,2], collapse=", ")
            meanlist[[i]] <- c(meanlist[[i]], as.numeric(fc[match,][,2]))
          }
          result[i,j+3] = fold #Saves the fold change to correct column
          i = i+1
          } else {
            i = 1+1
          }
      }
      j = j+1
  }

  #calculate mean fold change and standard deviation of each gene
  for(i in 1:length(glist[ ,1])) {
    mn <- mean(meanlist[[i]], na.rm = TRUE)
    stdev <- sd(meanlist[[i]], na.rm = TRUE)
    result[i,2] <- mn
    result[i,3] <- stdev

  }
  result
}
```

H Lists of the enriched GO-terms

Table H1: Enriched terms of A1 data subset consisting of cells from peripheral blood, GO category biological process. The cutoff for P-value was 0.00001

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0002429	0.000	4.825	8	31	180	immune response-activating cell surface receptor signaling pathway
GO:0050853	0.000	18.336	1	13	29	B cell receptor signaling pathway
GO:0002253	0.000	3.237	15	42	346	activation of immune response
GO:0051249	0.000	3.533	12	36	273	regulation of lymphocyte activation
GO:0002764	0.000	3.002	17	45	397	immune response-regulating signaling pathway
GO:0002696	0.000	3.916	9	31	214	positive regulation of leukocyte activation
GO:0042113	0.000	4.222	8	27	174	B cell activation
GO:0050865	0.000	3.015	15	39	340	regulation of cell activation
GO:0030888	0.000	8.876	2	13	46	regulation of B cell proliferation
GO:0046651	0.000	3.802	8	27	190	lymphocyte proliferation
GO:0006955	0.000	2.025	49	86	1115	immune response
GO:0070661	0.000	3.578	9	27	200	leukocyte proliferation
GO:0050854	0.000	11.184	1	9	27	regulation of antigen receptor-mediated signaling pathway
GO:0050852	0.000	4.632	4	16	94	T cell receptor signaling pathway
GO:0002252	0.000	2.300	21	44	488	immune effector process
GO:0050870	0.000	3.532	7	21	156	positive regulation of T cell activation
GO:0048518	0.000	1.557	141	186	3246	positive regulation of biological process
GO:0002682	0.000	1.894	39	66	902	regulation of immune system process
GO:0048583	0.000	1.587	101	140	2319	regulation of response to stimulus

Table H2: Enriched terms of A1 data subset consisting of cells from peripheral blood, GO categorymolecular function. The cutoff for P-value was 0.005

GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0044822	0.000	1.714	60	94	914	poly(A) RNA binding
GO:0003714	0.000	2.629	11	26	169	transcription corepressor activity
GO:0000989	0.000	1.760	29	48	445	transcription factor binding transcription factor activity
GO:0003677	0.001	1.392	106	137	1607	DNA binding
GO:0003676	0.001	1.491	67	92	1065	nucleic acid binding
GO:0003823	0.001	3.300	4	12	64	antigen binding
GO:0003682	0.001	1.875	20	35	305	chromatin binding
GO:0032395	0.001	14.207	1	4	8	MHC class II receptor activity
GO:0004896	0.002	2.860	5	13	78	cytokine receptor activity
GO:0070577	0.002	7.109	1	5	15	histone acetyl-lysine binding
GO:0042393	0.002	2.750	5	13	81	histone binding
GO:0043548	0.004	5.923	1	5	17	phosphatidylinositol 3-kinase binding
GO:0016303	0.004	8.116	1	4	11	1-phosphatidylinositol-3-kinase activity
GO:0004906	0.004	Inf	0	2	2	interferon-gamma receptor activity
GO:0034046	0.004	Inf	0	2	2	poly(G) binding
GO:0042610	0.004	Inf	0	2	2	CD8 receptor binding
GO:0003727	0.005	3.454	3	8	41	single-stranded RNA binding

Table H3: Enriched terms of A2 data subset consisting of cells from bone marrow, GO category biological process. The cutoff for P-value was 0.000001

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0002684	0.000	3.613	17	51	580	positive regulation of immune system process
GO:0002429	0.000	6.294	5	27	180	immune response-activating cell surface receptor signaling pathway
GO:0050853	0.000	24.248	1	12	29	B cell receptor signaling pathway
GO:0002253	0.000	3.922	10	34	346	activation of immune response
GO:0051251	0.000	5.139	6	25	197	positive regulation of lymphocyte activation
GO:0006955	0.000	2.514	33	69	1115	immune response
GO:0042113	0.000	5.359	5	23	174	B cell activation
GO:0002764	0.000	3.597	12	36	397	immune response-regulating signaling pathway
GO:0050867	0.000	4.647	7	26	224	positive regulation of cell activation
GO:0002694	0.000	3.940	9	31	312	regulation of leukocyte activation
GO:0006402	0.000	4.856	5	19	155	mRNA catabolic process
GO:0000184	0.000	6.156	3	15	99	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
GO:0045047	0.000	6.419	3	14	89	protein targeting to ER
GO:0019058	0.000	5.115	4	17	132	viral life cycle
GO:0030098	0.000	3.831	7	23	233	lymphocyte differentiation
GO:0042110	0.000	3.293	10	28	328	T cell activation
GO:0006414	0.000	6.557	2	13	81	translational elongation
GO:0030888	0.000	9.463	1	10	46	regulation of B cell proliferation
GO:0046651	0.000	4.090	6	20	190	lymphocyte proliferation
GO:0002252	0.000	2.751	14	35	488	immune effector process

Table H4: Enriched terms of A2 data subset consisting of cells from bone marrow, GO categorymolecular function. The cutoff for P-value was 0.005

GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0044822	0.000	2.440	44	91	833	poly(A) RNA binding
GO:0003735	0.000	5.623	6	25	107	structural constituent of ribosome
GO:0003676	0.000	1.565	142	192	2672	nucleic acid binding
GO:0019899	0.000	1.755	55	87	1030	enzyme binding
GO:0042393	0.000	3.355	5	15	96	histone binding
GO:0003779	0.001	2.053	16	30	297	actin binding
GO:0019901	0.001	1.974	18	33	339	protein kinase binding
GO:0030331	0.001	7.178	1	6	21	estrogen receptor binding
GO:0050733	0.001	26.812	0	3	5	RS domain binding
GO:0043548	0.002	7.466	1	5	17	phosphatidylinositol 3-kinase binding
GO:0008143	0.003	8.948	1	4	12	poly(A) binding
GO:0046625	0.003	8.948	1	4	12	sphingolipid binding
GO:0034988	0.003	Inf	0	2	2	Fc-gamma receptor I complex binding
GO:0042610	0.003	Inf	0	2	2	CD8 receptor binding
GO:0050072	0.003	Inf	0	2	2	m7G(5')pppN diphosphatase activity
GO:0003682	0.003	1.838	16	28	305	chromatin binding
GO:0003729	0.004	2.830	4	11	81	mRNA binding

Table H5: Enriched terms of A3 data subset consisting of cells from peripheral blood, GO category biological process. The cutoff for P-value was 0.0005

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0046649	0.000	3.026	8	22	174	lymphocyte activation
GO:0010829	0.000	40.317	0	4	6	negative regulation of glucose transport
GO:0070059	0.000	11.223	1	5	14	intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress
GO:0001775	0.000	2.223	13	27	281	cell activation
GO:0016574	0.000	10.099	1	5	15	histone ubiquitination

Table H6: Enriched terms of A3 data subset consisting of cells from peripheral blood, GO category molecular function. The cutoff for P-value was 0.01

GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0035091	0.000	4.104	3	11	94	phosphatidylinositol binding
GO:0016881	0.001	2.922	5	14	163	acid-amino acid ligase activity
GO:0004842	0.001	3.033	5	13	146	ubiquitin-protein ligase activity
GO:0043325	0.002	18.092	0	3	8	phosphatidylinositol-3,4-bisphosphate binding
GO:0004708	0.006	30.024	0	2	4	MAP kinase kinase activity

Table H7: Enriched terms of A4 data subset consisting of cells from bone marrow, GO category biological process. The cutoff for P-value was 0.001

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0045321	0.000	4.137	5	19	201	leukocyte activation
GO:0042113	0.000	7.077	2	10	64	B cell activation
GO:0045580	0.000	10.997	1	7	31	regulation of T cell differentiation
GO:0051252	0.000	1.985	38	61	1415	regulation of RNA metabolic process
GO:0010556	0.000	1.904	41	63	1517	regulation of macromolecule biosynthetic process
GO:0080090	0.000	1.812	56	79	2072	regulation of primary metabolic process
GO:0032774	0.000	1.863	41	63	1542	RNA biosynthetic process
GO:0030098	0.000	4.697	2	10	91	lymphocyte differentiation
GO:0051171	0.000	1.822	45	67	1687	regulation of nitrogen compound metabolic process
GO:0002764	0.000	3.370	5	14	174	immune response-regulating signaling pathway
GO:0090304	0.000	1.751	55	77	2053	nucleic acid metabolic process
GO:0006355	0.000	1.827	36	56	1361	regulation of transcription, DNA-templated
GO:0051249	0.000	4.174	3	10	101	regulation of lymphocyte activation
GO:0042110	0.000	3.811	3	11	121	T cell activation
GO:0002253	0.000	3.556	4	12	141	activation of immune response
GO:0031326	0.000	1.771	42	62	1570	regulation of cellular biosynthetic process
GO:0045727	0.000	9.319	1	5	25	positive regulation of translation
GO:0002684	0.001	2.933	6	15	212	positive regulation of immune system process
GO:0050865	0.001	3.642	3	11	126	regulation of cell activation
GO:1901362	0.001	1.711	46	65	1701	organic cyclic compound biosynthetic process
GO:0019438	0.001	1.715	45	64	1668	aromatic compound biosynthetic process
GO:0018130	0.001	1.713	45	64	1669	heterocycle biosynthetic process
GO:0039535	0.001	Inf	0	2	2	regulation of RIG-I signaling pathway
GO:0002683	0.001	4.567	2	8	74	negative regulation of immune system process
GO:0002429	0.001	4.498	2	8	75	immune response-activating cell surface receptorn signaling pathway
GO:0065007	0.001	1.695	103	123	3851	biological regulation

Table H8: Enriched terms of A4 data subset consisting of cells from bone marrow, GO category molecular function. The cutoff for P-value was 0.01

GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0003712	0.003	2.634	5	13	209	transcription cofactor activity
GO:0032266	0.003	12.812	0	3	12	phosphatidylinositol-3-phosphate binding
GO:0051059	0.003	12.812	0	3	12	NF-kappaB binding
GO:0005161	0.004	38.247	0	2	4	platelet-derived growth factor receptor binding
GO:0003823	0.005	10.479	0	3	14	antigen binding
GO:0000988	0.005	2.441	6	13	224	protein binding transcription factor activity
GO:0004716	0.006	25.494	0	2	5	receptor signaling protein tyrosine kinase activity
GO:0042393	0.007	4.500	1	5	48	histone binding
GO:0016796	0.009	8.230	0	3	17	exonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 5'-phosphomonoesters
GO:0042287	0.009	19.117	0	2	6	MHC protein binding
GO:0019900	0.010	2.416	5	11	190	kinase binding

Table H9: Enriched terms of CC1 data subset consisting of plasma cells grown for 20 days, GO category biological process. The cutoff for P-value was 0.00001

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0002253	0.000	5.239	5	23	385	activation of immune response
GO:0006955	0.000	3.073	17	43	1270	immune response
GO:0002429	0.000	6.724	3	15	193	immune response-activating cell surface receptor signaling pathway
GO:0002764	0.000	4.276	6	22	442	immune response-regulating signaling pathway
GO:0002684	0.000	3.628	9	27	644	positive regulation of immune system process
GO:0042113	0.000	6.136	3	14	195	B cell activation
GO:0045321	0.000	3.553	8	25	603	leukocyte activation
GO:0050853	0.000	15.878	0	6	35	B cell receptor signaling pathway
GO:0051251	0.000	5.044	3	13	216	positive regulation of lymphocyte activation
GO:0048583	0.000	2.043	37	62	2767	regulation of response to stimulus

Table H10: Enriched terms of CC1 data subset consisting of plasma cells grown for 20 days, GO category molecular function. The cutoff for P-value was 0.001

GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0019899	0.000	2.136	34	65	1195	enzyme binding
GO:0003735	0.000	4.680	4	17	145	structural constituent of ribosome
GO:0044822	0.000	2.099	32	60	1113	poly(A) RNA binding
GO:0019901	0.000	2.421	11	25	390	protein kinase binding
GO:0005488	0.001	1.572	343	369	12044	binding

Table H11: Enriched terms of CC2 data subset consisting of plasma cells grown for 41 days, GO category biological process. The cutoff for P-value was 0.00002

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0001775	0.000	4.159	10	33	816	cell activation
GO:0006955	0.000	3.270	15	40	1270	immune response
GO:0042113	0.000	5.864	2	12	195	B cell activation
GO:0023051	0.000	2.213	28	51	2371	regulation of signaling
GO:0046649	0.000	4.464	3	14	315	lymphocyte activation
GO:0048583	0.000	2.100	32	56	2767	regulation of response to stimulus
GO:0010646	0.000	2.143	28	50	2378	regulation of cell communication
GO:0002429	0.000	5.372	2	11	193	immune response-activating cell surface receptor signaling pathway
GO:0007165	0.000	1.939	52	78	4426	signal transduction
GO:0002520	0.000	3.480	6	18	489	immune system development
GO:0002694	0.000	3.977	4	15	355	regulation of leukocyte activation

Table H12: Enriched terms of CC2 data subset consisting of plasma cells grown for 41 days, GO category molecular function. The cutoff for P-value was 0.005

GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0019899	0.000	2.304	19	40	1195	enzyme binding
GO:0019904	0.000	2.548	9	21	546	protein domain specific binding
GO:0004715	0.001	7.684	1	5	45	non-membrane spanning protein tyrosine kinase activity
GO:0019901	0.001	2.571	6	15	384	protein kinase binding
GO:0051434	0.002	60.868	0	2	4	BH3 domain binding
GO:0016773	0.002	2.080	11	22	692	phosphotransferase activity, alcohol group as acceptor
GO:0016301	0.002	2.019	12	23	745	kinase activity
GO:0008432	0.004	30.430	0	2	6	JUN kinase binding
GO:0055103	0.004	30.430	0	2	6	ligase regulator activity
GO:0043548	0.005	10.176	0	3	21	phosphatidylinositol 3-kinase binding
GO:0003823	0.005	4.950	1	5	67	antigen binding
GO:0004674	0.005	2.260	7	15	430	protein serine/threonine kinase activity

Table H13: Enriched terms of CC3 data subset consisting of cells from bone marrow, GO category biological process. The cutoff for P-value was 0.000001

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0034976	0.000	11.682	2	18	130	response to endoplasmic reticulum stress
GO:0035966	0.000	10.070	2	17	139	response to topologically incorrect protein
GO:0034620	0.000	13.699	1	14	87	cellular response to unfolded protein
GO:0006984	0.000	12.040	1	14	97	ER-nucleus signaling pathway
GO:0006987	0.000	14.909	1	11	63	activation of signaling protein activity involved in unfolded protein response
GO:0032069	0.000	12.914	1	11	71	regulation of nuclease activity
GO:0018279	0.000	10.218	1	12	95	protein N-linked glycosylation via asparagine
GO:0070972	0.000	8.302	2	13	124	protein localization to endoplasmic reticulum
GO:0044267	0.000	2.205	50	85	3389	cellular protein metabolic process
GO:0002478	0.000	6.543	2	14	166	antigen processing and presentation of exogenous peptide antigen
GO:0019882	0.000	5.525	3	16	223	antigen processing and presentation
GO:0006488	0.000	17.950	0	7	34	dolichol-linked oligosaccharide biosynthetic process
GO:0030433	0.000	16.152	1	7	37	ER-associated ubiquitin-dependent protein catabolic process

Table H14: Enriched terms of CC3 data subset consisting of cells from bone marrow, GO category molecular function. The cutoff for P-value was 0.0001

GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0004579	0.000	140.560	0	7	8	dolichyl-diphosphooligosaccharide-protein glycotransferase activity
GO:0003954	0.000	9.240	2	11	35	NADH dehydrogenase activity
GO:0008137	0.000	9.240	2	11	35	NADH dehydrogenase (ubiquinone) activity
GO:0016758	0.000	3.203	9	25	184	transferase activity, transferring hexosyl groups
GO:0032395	0.000	24.055	1	6	11	MHC class II receptor activity
GO:0016655	0.000	5.830	2	11	49	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor
GO:0016491	0.000	1.890	33	57	680	oxidoreductase activity
GO:0003756	0.000	12.024	1	6	16	protein disulfide isomerase activity

Table H15: Enriched terms of C1 data subset consisting of plasmablasts, GO category biological process. The cutoff for P-value was 0.001

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0002684	0.000	2.516	9	21	631	positive regulation of immune system process
GO:0036092	0.000	29.784	0	3	10	phosphatidylinositol-3-phosphate biosynthetic process
GO:0045017	0.000	3.756	3	11	220	glycerolipid biosynthetic process
GO:0007067	0.000	3.058	5	14	343	mitosis
GO:0030071	0.001	8.336	1	5	47	regulation of mitotic metaphase/anaphase transition
GO:0007079	0.001	138.352	0	2	3	mitotic chromosome movement towards spindle pole
GO:0006955	0.001	1.998	18	32	1224	immune response
GO:0030888	0.001	7.291	1	5	53	regulation of B cell proliferation

Table H16: Enriched terms of C1 data subset consisting of plasmablasts, GO category molecular function. The cutoff for P-value was 0.005

GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0043813	0.000	Inf	0	2	2	phosphatidylinositol-3,5-bisphosphate 5-phosphatase activity
GO:0034596	0.001	70.890	0	2	4	phosphatidylinositol phosphate 4-phosphatase activity
GO:0000099	0.003	35.440	0	2	6	sulfur amino acid transmembrane transporter activity
GO:0015175	0.003	11.861	0	3	21	neutral amino acid transmembrane transporter activity
GO:0019899	0.003	1.857	16	28	1159	enzyme binding
GO:0016308	0.004	28.350	0	2	7	1-phosphatidylinositol-4-phosphate 5-kinase activity
GO:0043014	0.004	10.164	0	3	24	alpha-tubulin binding

Table H17: Enriched terms of GT1 data subset consisting of cells from spleen, GO category biological process. The cutoff for P-value was 0.0001

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006955	0.000	2.738	15	35	1115	immune response
GO:0051251	0.000	5.466	3	13	197	positive regulation of lymphocyte activation
GO:0042110	0.000	4.008	5	16	328	T cell activation
GO:0050867	0.000	4.754	3	13	224	positive regulation of cell activation
GO:0031295	0.000	10.097	1	7	59	T cell costimulation
GO:0002429	0.000	4.969	2	11	180	immune response-activating cell surface receptor signaling pathway
GO:0060333	0.000	8.600	1	7	68	interferon-gamma-mediated signaling pathway
GO:0002694	0.000	3.622	4	14	312	regulation of leukocyte activation

Table H18: Enriched terms of GT1 data subset consisting of cells from spleen, GO category molecular function. The cutoff for P-value was 0.01

GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0003823	0.000	12.684	1	9	64	antigen binding
GO:0032395	0.000	44.765	0	3	8	MHC class II receptor activity
GO:0032403	0.002	2.358	8	17	582	protein complex binding
GO:0023026	0.006	21.165	0	2	9	MHC class II protein complex binding
GO:0003746	0.009	16.458	0	2	11	translation elongation factor activity

I Differentially expressed genes

Table I1: The most differentially expressed genes of subset A1. *P*-value cutoff was 1^{-8}

B-cell translocation gene 1, anti-proliferative
B-cell translocation gene 1, anti-proliferative
actin binding LIM protein 1
major histocompatibility complex, class II, DP beta 1
ribosome binding protein 1
synaptophysin-like 1
kelch-like ECH-associated protein 1
trafficking protein particle complex 12
nucleobindin 2
phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit delta
major histocompatibility complex, class II, DM beta
cyclin-dependent kinase 5
hematopoietically expressed homeobox
lymphotoxin beta (TNF superfamily, member 3)
TSC22 domain family, member 3
Wolf-Hirschhorn syndrome candidate 1
UDP-N-acteylglucosamine pyrophosphorylase 1
T-cell leukemia/lymphoma 1A
chemokine (C-X-C motif) receptor 4
ER membrane protein complex subunit 1
septin 7
tripartite motif containing 22
elongation factor, RNA polymerase II, 2
CD72 molecule
hematopoietically expressed homeobox
membrane-spanning 4-domains, subfamily A, member 1
5'-nucleotidase domain containing 2
phosphopantothenoylcysteine decarboxylase
SLAM family member 7
ninjurin 2
aquaporin 3 (Gill blood group)
T-cell leukemia/lymphoma 1A

Table I2: The most differentially expressed genes of subset A2. P-value cutoff was 1^{-8}

KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 2
cytoskeleton-associated protein 4
signal sequence receptor, delta
acidic (leucine-rich) nuclear phosphoprotein 32 family, member A
major histocompatibility complex, class II, DP beta 1
peroxiredoxin 4
peptidylglycine alpha-amidating monooxygenase
hematopoietic cell-specific Lyn substrate 1
metastasis suppressor 1
trafficking protein particle complex 12
TIMP metallopeptidase inhibitor 2
interleukin 4 receptor
nucleobindin 2
inositol 1,4,5-trisphosphate receptor, type 1
phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit delta
major histocompatibility complex, class II, DM beta
interferon regulatory factor 8
CD37 molecule
selectin L
CD22 molecule
interferon (alpha, beta and omega) receptor 2
CD19 molecule
lymphocyte antigen 96
family with sequence similarity 65, member B
lymphotoxin beta (TNF superfamily, member 3)
RAS guanyl releasing protein 2 (calcium and DAG-regulated)
tumor necrosis factor, alpha-induced protein 8
DnaJ (Hsp40) homolog, subfamily C, member 3
dolichyl-diphosphooligosaccharide–protein glycosyltransferase subunit (non-catalytic)
major histocompatibility complex, class II, DR alpha
transmembrane 9 superfamily member 1
UDP-N-acteylglucosamine pyrophosphorylase 1
phospholipase A2, group XVI
SEPT5-GP1BB readthrough
CD69 molecule
family with sequence similarity 65, member B
membrane-spanning 4-domains, subfamily A, member 1
sorting nexin 3
major histocompatibility complex, class II, DR alpha
immunoglobulin lambda constant 1 (Mcg marker)
lamin A/C
protein tyrosine phosphatase, receptor type, C
protein tyrosine phosphatase, receptor type, C
immunoglobulin heavy constant mu
tripartite motif containing 22
insulin receptor
nucleosome assembly protein 1-like 1
elongation factor, RNA polymerase II, 2
CD72 molecule
hematopoietically expressed homeobox
immunoglobulin lambda-like polypeptide 3, pseudogene
ring finger protein 5, E3 ubiquitin protein ligase
microtubule associated serine/threonine kinase 1
membrane-spanning 4-domains, subfamily A, member 1
major histocompatibility complex, class II, DM alpha
nicotinamide phosphoribosyltransferase
proline-rich nuclear receptor coactivator 2
ubiquitin-conjugating enzyme E2, J1
5'-nucleotidase domain containing 2
signal recognition particle receptor, B subunit
ubiquitin-like modifier activating enzyme 5
SIL1 nucleotide exchange factor
FK506 binding protein 11, 19 kDa
SLAM family member 7
BTB and CNC homology 1, basic leucine zipper transcription factor 2
transmembrane protein 208
Fas apoptotic inhibitory molecule 3
Fas apoptotic inhibitory molecule 3
chromosome 19 open reading frame 10
capping protein (actin filament) muscle Z-line, beta
CD22 molecule
aquaporin 3 (Gill blood group)

Table I3: The most differentially expressed genes of subset A3. *P*-value cutoff was 1^{-7}

peptidylprolyl isomerase (cyclophilin)-like 1
DnaJ (Hsp40) homolog, subfamily C, member 1
solute carrier family 35 (GDP-fucose transporter), member C1
B-cell CLL/lymphoma 11A (zinc finger protein)
KIAA1147
v-ets avian erythroblastosis virus E26 oncogene homolog 1
forkhead box P1
forkhead box P1
homeodomain interacting protein kinase 2
pleckstrin homology domain containing, family A (phosphoinositide binding specific) member 2
TSEN15 tRNA splicing endonuclease subunit
phosphoprotein associated with glycosphingolipid microdomains 1
ankyrin repeat domain 28
zinc finger, CCHC domain containing 7
La ribonucleoprotein domain family, member 1B
hydrogen voltage-gated channel 1
centromere protein W
AF4/FMR2 family, member 3
membrane-spanning 4-domains, subfamily A, member 1
membrane-spanning 4-domains, subfamily A, member 1
membrane-associated ring finger (C3HC4) 1, E3 ubiquitin protein ligase
family with sequence similarity 26, member F
T-cell activation RhoGTPase activating protein
long intergenic non-protein coding RNA 926
baculoviral IAP repeat containing 3
Fc receptor-like 3
immunoglobulin lambda variable 1-44
tetratricopeptide repeat domain 39C
Fc receptor-like 1

Table I4: The most differentially expressed genes of subset A4. P-value cutoff was 1^{-7}

signal sequence receptor, gamma (translocon-associated protein gamma)
peptidylprolyl isomerase (cyclophilin)-like 1
DnaJ (Hsp40) homolog, subfamily C, member 1
DnaJ (Hsp40) homolog, subfamily C, member 1
regulatory factor X, 7
fucosidase, alpha-L- 2, plasma
VCP-interacting membrane protein
forkhead box P1
dystrobrevin binding protein 1
kidney associated antigen 1
cytochrome b561 family, member A3
v-ets avian erythroblastosis virus E26 oncogene homolog 1
forkhead box P1
forkhead box P1
scavenger receptor class B, member 2
pleckstrin homology domain containing, family A (phosphoinositide binding specific) member 2
collagen triple helix repeat containing 1
v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog K
zinc finger, CCHC domain containing 7
major histocompatibility complex, class II, DO alpha
hydrogen voltage-gated channel 1
elongation factor, RNA polymerase II, 2
AF4/FMR2 family, member 3
transmembrane anterior posterior transformation 1
protein kinase C, beta
napsin B aspartic peptidase, pseudogene
calcium/calmodulin-dependent protein kinase II delta
membrane-spanning 4-domains, subfamily A, member 1
ubiquitin-conjugating enzyme E2, J1
ITGB2 antisense RNA 1
ankyrin repeat domain 28
ribosomal modification protein rimK-like family member B
membrane-associated ring finger (C3HC4) 1, E3 ubiquitin protein ligase
family with sequence similarity 26, member F
DENN/MADD domain containing 6B
forkhead box P1
meiosis inhibitor 1
long intergenic non-protein coding RNA 926
family with sequence similarity 129, member C
mucin 17, cell surface associated
myosin, heavy chain 7B, cardiac muscle, beta
immunoglobulin lambda variable cluster
NLR family, CARD domain containing 3
CASP8 and FADD-like apoptosis regulator
CD53 molecule
Fc receptor-like 1

Table I5: The most differentially expressed genes of subset CC1. P-value cutoff was 1^{-5}

phosphoinositide-3-kinase adaptor protein 1
Fc receptor-like 2
von Willebrand factor A domain containing 5A
coatomer protein complex, subunit beta 2 (beta prime)
FYN oncogene related to SRC, FGR, YES
chemokine (C-C motif) receptor 6
quinolinate phosphoribosyltransferase
ALG3, alpha-1,3- mannosyltransferase
met proto-oncogene
apolipoprotein O
nucleobindin 1
translocation associated membrane protein 1
eukaryotic translation initiation factor 3, subunit D
kynureninase
lipocalin-like 1
DnaJ (Hsp40) homolog, subfamily B, member 11
hematopoietically expressed homeobox
neurensin 2
thioredoxin domain containing 11
Fas apoptotic inhibitory molecule 3
baculoviral IAP repeat containing 3
T-cell leukemia/lymphoma 1A
proteolipid protein 1
androgen-induced 1
psoriasis susceptibility 1 candidate 1
adenosine deaminase
lectin, galactoside-binding, soluble, 3
ankyrin repeat and SOCS box containing 16
non-POU domain containing, octamer-binding
SEC24 family member A
regulator of chromosome condensation (RCC1) and BTB (POZ) domain containing protein 2
lymphoblastic leukemia associated hematopoiesis regulator 1
glutamine–fructose-6-phosphate transaminase 1
CD38 molecule
integrin, alpha 6
Sec23 homolog B (S. cerevisiae)
Sec61 gamma subunit
lymphotoxin beta (TNF superfamily, member 3)
ST6 beta-galactosamide alpha-2,6-sialyltranferase 1

Table I6: The most differentially expressed genes of subset CC2. P-value cutoff was 1^{-6}

major histocompatibility complex, class II, DO alpha
signal sequence receptor, delta
dipeptidase 2
growth arrest and DNA-damage-inducible, alpha
kynureninase
serglycin
hematopoietically expressed homeobox
tyrosyl-tRNA synthetase
pelota homolog (Drosophila)
chromosome 8 open reading frame 48
FK506 binding protein 11, 19 kDa
T-cell leukemia/lymphoma 1A
CDGSH iron sulfur domain 2
psoriasis susceptibility 1 candidate 1
ankyrin repeat and SOCS box containing 16
RAS guanyl releasing protein 2 (calcium and DAG-regulated)
lymphotoxin beta (TNF superfamily, member 3)
membrane-spanning 4-domains, subfamily A, member 1
PR domain containing 1, with ZNF domain

Table I7: The most differentially expressed genes of subset CC3. P-value cutoff was 1^{-6}

sphingosine-1-phosphate receptor 1
major histocompatibility complex, class II, DO alpha
B-cell scaffold protein with ankyrin repeats 1
RIC8 guanine nucleotide exchange factor B
family with sequence similarity 129, member C
BTB and CNC homology 1, basic leucine zipper transcription factor 2
lysosomal-associated membrane protein 2
FK506 binding protein 2, 13kDa
GRB2-binding adaptor protein, transmembrane
phosphohistidine phosphatase 1
signal sequence receptor, delta
SMAD family member 3
FYN oncogene related to SRC, FGR, YES
v-ets avian erythroblastosis virus E26 oncogene homolog 1
chemokine (C-C motif) receptor 6
small integral membrane protein 7
chemokine (C-C motif) receptor 7
histone cluster 1, H2bg
MOB kinase activator 3B
CD72 molecule
ADP-ribosylation factor 4
family with sequence similarity 65, member B
TNF receptor-associated factor 5
kynureninase
hematopoietically expressed homeobox
protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 2
KIAA0226-like
GDP-mannose pyrophosphorylase B
heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa)
AF4/FMR2 family, member 3
membrane-spanning 4-domains, subfamily A, member 1
tripartite motif containing 22
WD repeat domain, phosphoinositide interacting 1
chromosome 19 open reading frame 10
T-cell leukemia/lymphoma 1A
G protein-coupled receptor 132
methyltransferase like 1
BTB and CNC homology 1, basic leucine zipper transcription factor 2
metastasis suppressor 1
membrane-associated ring finger (C3HC4) 1, E3 ubiquitin protein ligase
napsin B aspartic peptidase, pseudogene
mesencephalic astrocyte-derived neurotrophic factor
CD38 molecule
chemokine (C-X-C motif) receptor 5
ATP-binding cassette, sub-family B (MDR/TAP), member 9
RAS guanyl releasing protein 2 (calcium and DAG-regulated)
lymphotoxin beta (TNF superfamily, member 3)
lymphotoxin beta (TNF superfamily, member 3)
CD247 molecule
membrane-spanning 4-domains, subfamily A, member 1
guanine nucleotide binding protein (G protein), beta 5
sphingosine-1-phosphate receptor 1
transmembrane protein 208
immunoglobulin kappa variable 1/ORY-1 (pseudogene)
signal peptidase complex subunit 3 homolog (S. cerevisiae)

Table I8: The most differentially expressed genes of subset C1. P-value cutoff was 1^{-6}

decapping enzyme, scavenger
signal transducer and activator of transcription 6, interleukin-4 induced
chromosome 12 open reading frame 42
T-cell leukemia/lymphoma 1A
phosphopantothenoylcysteine decarboxylase
methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase
apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3B
fibronectin type III domain containing 3B
selenoprotein K
guanine nucleotide binding protein (G protein), beta polypeptide 4
leucine aminopeptidase 3
SEC24 family member D
SEC24 family member A
PR domain containing 1, with ZNF domain
phosphoserine aminotransferase 1
solute carrier family 44 (choline transporter), member 1
heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa)

Table I9: The most differentially expressed genes of subset GT1. P-value cutoff was 1^{-4}

signal sequence receptor, alpha
cytochrome c-1
SWAP switching B-cell complex 70kDa subunit
leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 4
major histocompatibility complex, class II, DP alpha 1
immunoglobulin kappa constant
dystonin
zinc finger protein 419
SH3-domain binding protein 4
ribokinase